PASW® Decision Trees (Entscheidungsbäume) 18

Weitere Informationen zu SPSS Inc.-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

SPSS ist eine eingetragene Marke.

PASW ist eine eingetragene Marke von SPSS Inc..

Die SOFTWARE und die Dokumentation werden mit BESCHRÄNKTEN RECHTEN zur Verfügung gestellt. Verwendung, Vervielfältigung und Veröffentlichung durch die Regierung unterliegen den Beschränkungen in Unterabschnitt (c)(1)(ii) von The Rights in Technical Data and Computer Software unter 52.227-7013. Vertragspartner/Hersteller ist SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.
Patentnr. 7.023.453

Allgemeiner Hinweis: Andere in diesem Dokument verwendete Produktnamen werden nur zu Identifikationszwecken genannt und können Marken der entsprechenden Unternehmen sein.

Windows ist eine eingetragene Marke der Microsoft Corporation.

Apple, Mac und das Mac-Logo sind Marken von Apple Computer, Inc., die in den USA und in anderen Ländern eingetragen sind.

Dieses Produkt verwendet WinWrap Basic, Copyright 1993–2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Kein Teil dieser Publikation darf ohne vorherige Einwilligung des Herstellers in irgendeiner Form elektronisch, mechanisch, durch Fotokopieren, Aufzeichnen oder sonstige Mittel reproduziert, in einem Datenabfragesystem gespeichert oder übertragen werden.

Vorwort

PASW Statistics 18 ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Decision Trees (Entscheidungsbäume) bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Decision Trees (Entscheidungsbäume) müssen zusammen mit PASW Statistics 18 Core verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von Decision Trees (Entscheidungsbäume) Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS Inc. erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von Decision Trees (Entscheidungsbäume) Erweiterungsmodul.

Kompatibilität

PASW Statistics kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS Inc.. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS Inc. wenden, um Informationen zum Kundendienst, zu Zahlungen oder zu Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Core-System ausgeliefert.

Kundendienst

Wenden Sie sich bei Fragen zur Lieferung oder Ihrem Kundenkonto an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS Inc. bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Seminaren wünschen, wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit PASW Statistics oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen zum Technischen Support finden Sie auf der Website unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Das Handbuch *SPSS Statistics Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für PASW Statistics 18 ist geplant. Das Handbuch *SPSS Statistics Advanced Statistical Procedures Companion*, bei dem auch PASW Statistics 18 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Statistics Guide to Data Analysis* für PASW Statistics 18 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Inhalt

Teil I: Benutzerhandbuch

1 Erstellen von Entscheidungsbäumen 1

Auswählen von Kategorien 6
Validierung 7
Kriterien für den Aufbau des Baums 8
- Aufbaubegrenzungen 9
- CHAID-Kriterien 10
- CRT-Kriterien 12
- QUEST-Kriterien 14
- Beschneiden von Bäumen 15
- Surrogate 16

Optionen 16
- Fehlklassifizierungskosten 17
- Profite 18
- A-priori-Wahrscheinlichkeit 20
- Werte 21
- Fehlende Werte definieren 23

Speichern der Modelldaten 24
Ausgabe 25
- Baumanzeige 26
- Statistik 28
- Diagramme 32
- Auswahl- und Bewertungsregeln 38

2 Baumeditor 41

Arbeiten mit umfangreichen Bäumen 43
- Baumstruktur 43
- Skalieren der Baumanzeige 44
- Knotenübersichtsfenster 44

Steuern der im Baum angezeigten Daten 46
Ändern der Farben und Schriftarten im Baum 46

Regeln für die Auswahl oder Bewertung von Fällen 49
Filtern von Fällen 49
Speichern von Auswahl- und Bewertungsregeln 49

Teil II: Beispiele

3 Datenannahmen und -anforderungen 53

Auswirkungen des Messniveaus auf Baummodelle. 53
Dauerhafte Zuweisung des Messniveaus. 56
Auswirkungen der Wertelabels auf Baummodelle. 57
Zuweisen von Wertelabels zu allen Werten 58

4 Verwenden von Entscheidungsbäumen zur Bewertung des Kreditrisikos 60

Erstellen des Modells 60
Erstellen des CHAID-Baummodells 60
Auswahl der Zielkategorien 61
Angeben von Aufbaukriterien für Bäume 62
Auswahl zusätzlicher Ausgaben 63
Speichern vorhergesagter Werte. 65
Bewertung des Modells 66
Modellzusammenfassungstabelle 67
Baumdiagramm 68
Baumtabelle 69
Gewinne für Knoten 70
Gewinndiagramm 71
Indexdiagramm 72
Risikoschätzer und Klassifizierung 72
Vorhergesagte Werte 73
Verfeinern des Modells. 74
Auswählen der Fälle in Knoten. 74
Untersuchung der ausgewählten Fälle. 76
Zuweisen von Kosten zu den Ergebnissen 78
Zusammenfassung 82

5 Konstruieren eines Bewertungsmodells 83

Konstruieren des Modells 83
Bewertung des Modells 85
Modellübersicht 86
Baummodelldiagramm 87
Risikoschätzer 88
Anwenden des Modells auf eine andere Datendatei 89
Zusammenfassung 92

6 Fehlende Werte in Baummodellen 93

Fehlende Werte bei CHAID 94
CHAID-Ergebnisse 96
Fehlende Werte bei CRT 97
CRT-Ergebnisse 100
Zusammenfassung 102

Anhang

A Beispieldateien 103

Index 115

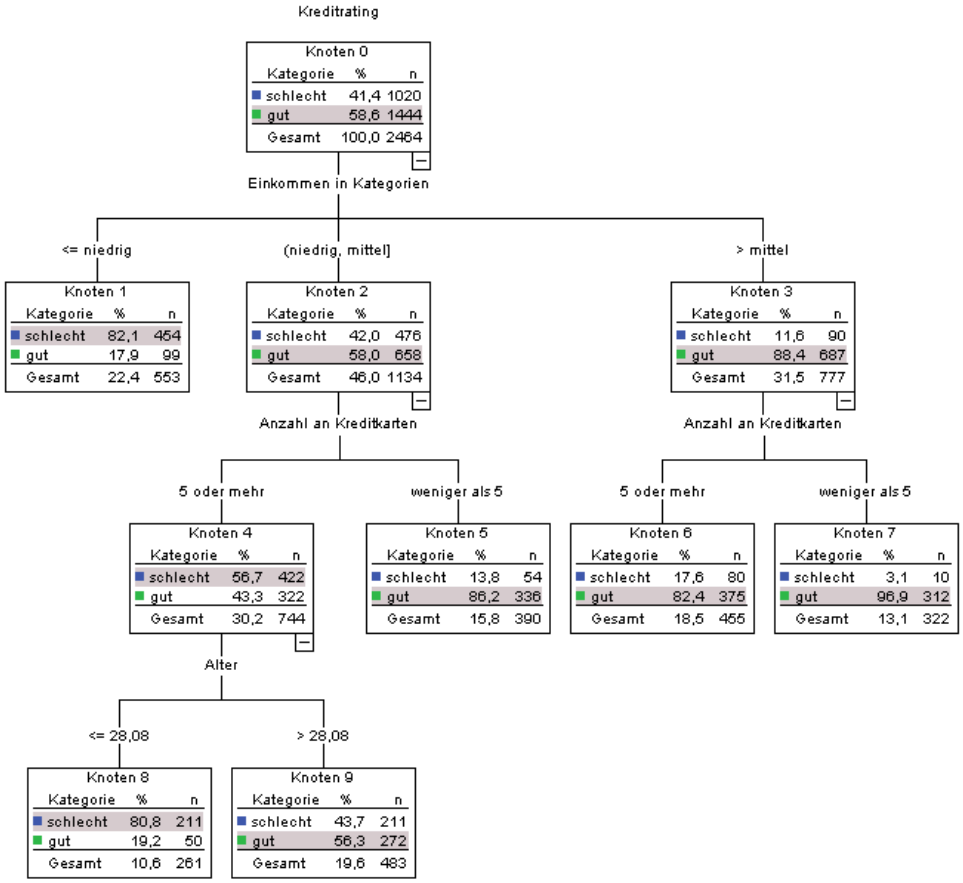
Teil I: Benutzerhandbuch

Kapitel

1

Erstellen von Entscheidungsbäumen

Abbildung 1-1
Entscheidungsbaum



Mit der Prozedur “Entscheidungsbaum” wird ein baumbasiertes Klassifizierungsmodell erstellt. Die Fälle werden in Gruppen klassifiziert oder es werden Werte für eine abhängige Variable (Zielvariable) auf der Grundlage der Werte von unabhängigen Variablen (Einflussvariablen) vorhergesagt. Die Prozedur umfasst Validierungswerkzeuge für die explorative und die bestätigende Klassifikationsanalyse.

Die Prozedur eignet sich für folgende Situationen:

Segmentierung. Ermitteln Sie Personen, die wahrscheinlich zu einer bestimmten Gruppe gehören.

Schichtung. Weisen Sie Fälle zu einer von mehreren Kategorien zu, z. B. Gruppen mit hohem, mittlerem oder niedrigem Risiko.

Vorhersage. Erstellen Sie Regeln und lassen Sie damit zukünftige Ereignisse voraussagen, z. B. die Wahrscheinlichkeit, dass eine Person mit dem Darlehen in Bezug gerät, oder den potenziellen Wiederverkaufswert eines Autos oder Hauses.

Dimensionsreduktion und Variablen-Screening. Wählen Sie eine geeignete Untergruppe an Einflussgrößen aus einer Vielzahl von Variablen aus und bauen Sie damit ein formales parametrisches Modell auf.

Erkennen von Wechselwirkungen. Ermitteln Sie Beziehungen, die nur für bestimmte Untergruppen gelten, und halten Sie diese in einem formalen parametrischen Modell fest.

Zusammenführung von Kategorien und Diskretisierung stetiger Variablen. Nehmen Sie die Umkodierung der Einflussgrößenkategorien und der stetigen Variablen bei minimalem Datenverlust vor.

Beispiel. Eine Bank möchte die Kreditantragsteller danach kategorisieren, ob sie ein annehmbares Kreditrisiko darstellen oder nicht. Auf der Grundlage verschiedener Faktoren (z. B. bekanntes Kreditrating bisheriger Kunden) können Sie ein Modell aufbauen, mit dem Sie vorhersagen, ob zukünftige Kunden mit ihren Darlehen in Verzug geraten würden.

Eine baumbasierte Analyse bietet einige ansprechende Möglichkeiten:

- Sie können homogene Gruppen mit hohem oder niedrigem Risiko erkennen.
- Regeln für Vorhersagen zu individuellen Fällen können leichter aufgestellt werden.

Erläuterung der Daten

Daten. Die abhängigen und die unabhängigen Variablen können wie folgt gestaltet sein:

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Metrisch.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

Häufigkeitsgewichtungen Wenn die Gewichtung aktiv ist, werden die Häufigkeitsgewichtungen auf die nächstliegende Ganzzahl gerundet. Fälle mit einer Gewichtung unter 0,5 erhalten einen Gewichtungswert von 0 und werden daher aus der Analyse ausgeschlossen.

Annahmen. Bei dieser Prozedur wird angenommen, dass allen Analysevariablen das entsprechende Messniveau zugewiesen wurde. Bei einigen Funktionen wird vorausgesetzt, dass ein Wertelabel für alle Werte der in der Analyse berücksichtigten abhängigen Variablen definiert wurde.

- **Messniveau.** Das Messniveau beeinflusst die Baumberechnungen. Sämtlichen Variablen sollte daher das geeignete Messniveau zugewiesen werden. Standardmäßig wird angenommen, dass numerische Variablen metrisch und String-Variablen nominal sind; dies spiegelt ggf. nicht das tatsächliche Messniveau wider. Der Variablentyp ist durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet.

Skalierung

Nominal

Ordinal

Sie können das Messniveau für eine Variable vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus.

- **Wertelabels.** In den Dialogfeldern für diese Prozedur wird angenommen, dass entweder alle der nichtfehlenden Werte einer kategorialen (nominalen, ordinalen) abhängigen Variablen über definierte Wertelabels verfügen oder keiner dieser Werte. Einige Funktionen sind nicht verfügbar, wenn nicht mindestens zwei nichtfehlende Werte der kategorialen abhängigen Variablen Wertelabels aufweisen. Wenn für mindestens zwei nichtfehlende Werte Wertelabels definiert sind, werden alle Fälle mit anderen Werten, die keine Wertelabels aufweisen, aus der Analyse ausgeschlossen.

So erhalten Sie Entscheidungsbäume

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Klassifizieren
    Baum...

Abbildung 1-2
Dialogfeld "Entscheidungsbaum"

- ▶ Wählen Sie eine abhängige Variable aus.
- ▶ Wählen Sie mindestens eine unabhängige Variable aus.
- ▶ Wählen Sie eine Aufbaumethode aus.

Die folgenden Optionen sind verfügbar:

- Ändern Sie das Messniveau für eine Variable in der Liste der Quellvariablen.
- Lassen Sie die erste Variable aus der Liste der unabhängigen Variablen als erste Teilungsvariable aufnehmen.
- Wählen Sie eine Einflussvariable aus, mit der definiert wird, wie viel Einfluss ein Fall auf den Aufbauprozess des Baums hat. Fälle mit niedrigeren Einflusswerten wirken sich weniger stark aus, Fälle mit höheren Werten entsprechend stärker. Die Einflussvariablen müssen positiv sein.
- Validieren Sie den Baum.
- Passen Sie die Kriterien für den Aufbau des Baums an.
- Speichern Sie die Endknotennummern, die vorhergesagten Werte und die vorhergesagten Wahrscheinlichkeiten als Variablen.
- Speichern Sie das Modell im XML-Format (PMML).

Ändern des Messniveaus

- ▶ Klicken Sie mit der rechten Maustaste auf eine Variable in der Liste der Quellvariablen.
- ▶ Wählen Sie ein Messniveau im Kontextmenü aus.

Das Messniveau wird vorübergehend für die Dauer der Prozedur "Entscheidungsbaum" geändert.

Aufbaumethoden

Die folgenden Aufbaumethoden sind verfügbar:

CHAID. Steht für "Chi-squared Automatic Interaction Detection", d. h. automatische Entdeckung von Zusammenhängen mittels Chi-Quadrat-Tests. In jedem Schritt bestimmt das CHAID-Verfahren diejenige unabhängige Variable (Einflussvariable/Prädiktor), die den stärksten Zusammenhang mit der abhängigen Variablen aufweist. Die Kategorien der einzelnen Einflussvariablen werden zusammengeführt, wenn sie im Hinblick auf die abhängige Variable nicht signifikant unterschiedlich sind.

Exhaustive CHAID. Eine Abwandlung von CHAID, die für jede Einflussvariable (Prädiktor) alle möglichen Aufteilungen untersucht.

CRT. Steht für Classification and Regression Trees, d. h. Klassifikations- und Regressionsbäume. CRT unterteilt die Daten in Segmente, die im Hinblick auf die abhängige Variable so homogen wie möglich sind. Ein Endknoten, in dem alle Fälle denselben Wert der abhängigen Variablen haben, ist ein homogener ("reiner") Knoten.

QUEST. Steht für Quick, Unbiased, Efficient Statistical Tree, d. h. schneller, unverzerrter, effizienter statistischer Baum. Dabei handelt es sich um ein schnelles Verfahren, das die in anderen Verfahren auftretende Verzerrung zugunsten von Prädiktoren (Einflussvariablen) mit vielen Kategorien vermeidet. QUEST kann nur dann gewählt werden, wenn die abhängige Variable nominal ist.

Jede Methode hat ihre Vorteile und Einschränkungen:

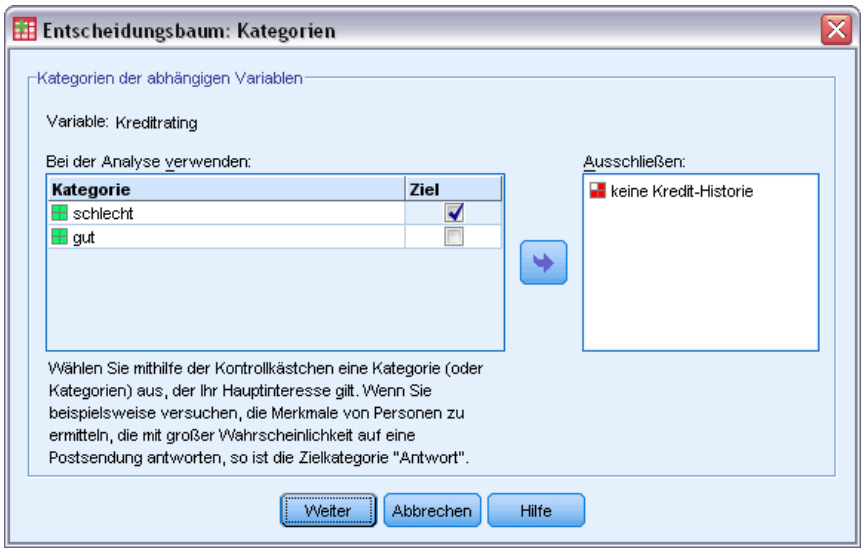
	CHAID*	**CRT**	**QUEST**
Chi-Quadrat-basiert**	O		
Surrogate für unabhängige Variablen (Einflussvariablen)		O	O
Beschneiden des Baums		O	O
Aufteilen mehrdimensionaler Knoten	O		
Aufteilen binärer Knoten		O	O
Einflussvariablen	O	O	
A-priori-Wahrscheinlichkeiten		O	O
Fehlklassifizierungskosten	O	O	O
Schnelle Berechnung	O		O

*Mit Exhaustive CHAID.

**Bei QUEST wird auch ein Chi-Quadrat-Maß für nominale unabhängige Variablen verwendet.

Auswählen von Kategorien

Abbildung 1-3
Dialogfeld "Kategorien"


Entscheidungsbaum: Kategorien
Kategorien der abhängigen Variablen
Variable: Kreditrating
Bei der Analyse verwenden:
Kategorie
Ziel
schlecht
gut
Ausschließen:
keine Kredit-Historie
Wählen Sie mithilfe der Kontrollkästchen eine Kategorie (oder Kategorien) aus, der Ihr Hauptinteresse gilt. Wenn Sie beispielsweise versuchen, die Merkmale von Personen zu ermitteln, die mit großer Wahrscheinlichkeit auf eine Postsendung antworten, so ist die Zielkategorie "Antwort".
Weiter
Abbrechen
Hilfe

Bei kategorialen (nominalen, ordinalen) abhängigen Variablen stehen folgende Möglichkeiten zur Auswahl:

- Kategorien festlegen, die im Diagramm angezeigt werden sollen.
- Relevante Zielkategorien auswählen

Kategorien ein-/ausschließen

Sie können die Analyse auf bestimmte Kategorien der abhängigen Variablen einschränken.

- Fälle mit Werten der abhängigen Variablen in der Liste "Ausschließen" werden bei der Analyse nicht berücksichtigt.
- Bei nominalen abhängigen Variablen können auch benutzerdefiniert fehlende Kategorien in die Analyse aufgenommen werden. (Standardmäßig werden benutzerdefiniert fehlende Kategorien in der Liste "Ausschließen" aufgeführt.)

Zielkategorien

Die ausgewählten (markierten) Kategorien werden als primär relevante Kategorien in der Analyse behandelt. Wenn Sie beispielsweise hauptsächlich die Personen ermitteln möchten, bei denen die Wahrscheinlichkeit groß ist, dass sie mit ihrem Darlehen in Verzug geraten, bestimmen Sie entsprechend die Kategorie für schlechtes Kreditrating als Zielkategorie.

- Es ist keine Standard-Zielkategorie festgelegt. Ist keine Kategorie ausgewählt, stehen einige Optionen für die Klassifikation sowie die Ausgabe im Zusammenhang mit dem Profit nicht zur Verfügung.

- Wenn mehrere Kategorien angegeben sind, werden separate Tabellen und Diagramme mit dem Profit in den einzelnen Zielkategorien erstellt.
- Die Kennzeichnung von einer oder mehreren Kategorien als Zielkategorien wirkt sich nicht auf das Baummodell, die Risikoschätzung und die Fehlklassifizierungsergebnisse aus.

"Kategorien" und Wertelabels

In diesem Dialogfeld sind definierte Wertelabels für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen ein Wertelabel besitzen.

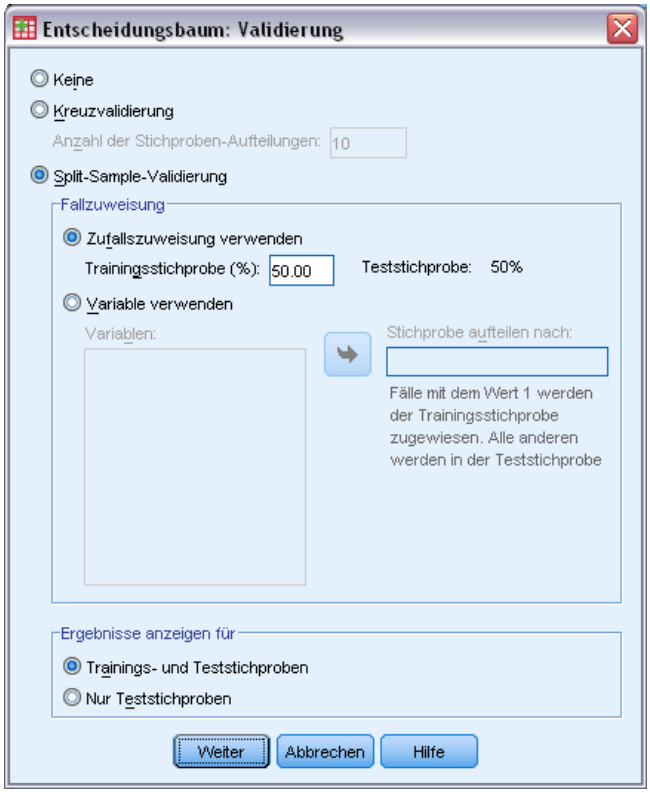
So können Sie Kategorien ein-/ausschließen und Zielkategorien auswählen:

▶ Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertelabels aus.

▶ Klicken Sie auf Kategorien.

Validierung

Abbildung 1-4
Dialogfeld "Validierung"



Mit der Validierung stellen Sie fest, wie gut sich die Baumstruktur auf eine größere Gesamtheit verallgemeinern lässt. Es stehen zwei Validierungsmethoden zur Auswahl: Kreuzvalidierung und Split-Sample-Validierung.

Kreuzvalidierung

Bei der Kreuzvalidierung wird die Stichprobe in mehrere Teilstichproben oder **Aufteilungen** gegliedert. Anschließend werden Baummodelle erzeugt; dabei werden nacheinander die Daten der einzelnen Stichproben ausgeschlossen. Der erste Baum beruht auf allen Fällen mit Ausnahme der Fälle in der ersten Stichprobenaufteilung, der zweite Baum auf allen Fällen mit Ausnahme der Fälle in der zweiten Stichprobenaufteilung usw. Bei jedem Baum wird jeweils das Fehlklassifizierungsrisiko geschätzt. Hierzu wird der Baum auf die Teilstichprobe angewendet, die beim Erstellen des Baums ausgeschlossen war.

- Sie können bis zu 25 Stichprobenaufteilungen angeben. Je höher der Wert, desto weniger Fälle werden in den einzelnen Baummodellen ausgeschlossen.
- Bei der Kreuzvalidierung entsteht ein einziges, endgültiges Baummodell. Die kreuzvalidierte Risikoschätzung für den fertigen Baum wird als Durchschnitt des Risikos bei allen Bäumen berechnet.

Split-Sample-Validierung

Bei der Split-Sample-Validierung wird das Modell mithilfe einer Trainingsstichprobe erzeugt und dann mit einer Teststichprobe überprüft.

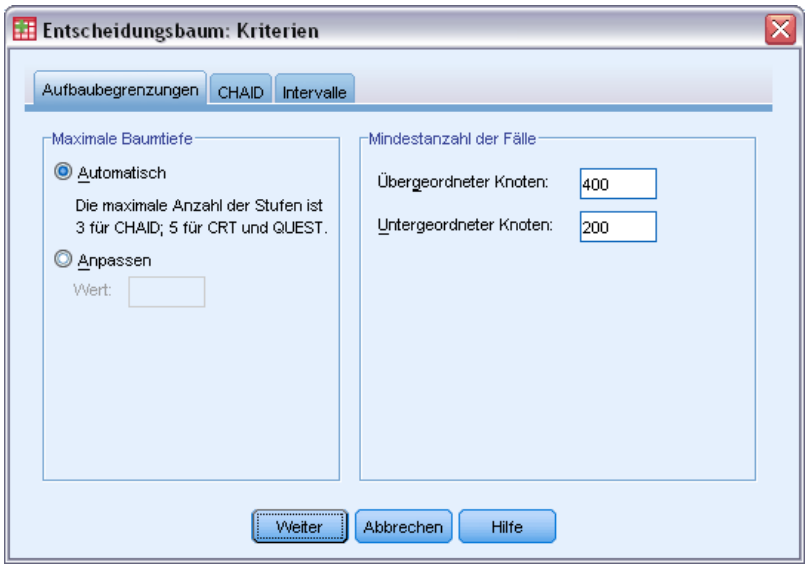
- Sie können eine Trainingsstichprobe angeben (als Prozentsatz der gesamten Stichprobengröße) oder auch eine Variable, mit der die Stichprobe in Trainings- und Teststichproben aufgeteilt wird.
- Wenn Sie die Trainings- und Teststichproben mithilfe einer Variablen festlegen, werden Fälle mit dem Wert 1 für die Variable in die Trainingsstichprobe übernommen, alle anderen Fälle in die Teststichprobe. Die abhängige Variable, die Gewichtungsvariable, die Einflussvariable sowie erzwungene unabhängige Variablen sind hier als Variable nicht zulässig.
- Die Ergebnisse können wahlweise für die Trainings- und Teststichproben oder auch nur für die Teststichprobe angezeigt werden.
- Bei kleinen Datendateien (Dateien mit nur wenigen Dateien) sollte die Split-Sample-Validierung nur nach sorgfältiger Erwägung verwendet werden. Kleine Trainingsstichproben können zu mangelhaften Modellen führen, weil einige Kategorien unter Umständen nicht genügend Fälle enthalten, damit der Baum ordnungsgemäß wachsen kann.

Kriterien für den Aufbau des Baums

Die verfügbaren Aufbaukriterien können von der Aufbaumethode und/oder dem Messniveau der abhängigen Variablen abhängen.

Aufbaubegrenzungen

Abbildung 1-5
Dialogfeld "Kriterien", Registerkarte "Aufbaubegrenzungen"



Auf der Registerkarte "Aufbaubegrenzungen" können Sie die Anzahl der Ebenen im Baum einschränken und die Mindestanzahl der Fälle für über- und untergeordnete Knoten steuern.

Maximale Baumtiefe. Steuert die maximale Anzahl der Aufbauebenen unterhalb des Stammknotens. Mit der Einstellung Automatisch wird der Baum auf drei (CHAID und Exhaustive CHAID) bzw. fünf Ebenen unterhalb des Stammknotens (CRT und QUEST) begrenzt.

Mindestanzahl der Fälle. Steuert die Mindestanzahl der Fälle für die Knoten. Knoten, die diese Kriterien nicht erfüllen, werden nicht aufgeteilt.

- Wenn Sie die Mindestwerte anheben, entstehen in der Regel Bäume mit weniger Knoten.
- Werden die Mindestwerte gesenkt, entstehen Bäume mit mehr Knoten.

Bei Datendateien mit nur wenigen Fällen führen die Standardwerte von 100 Fällen für übergeordnete Knoten und 50 Fällen für untergeordnete Knoten unter Umständen dazu, dass der resultierende Baum keine Knoten unterhalb des Stammknotens erhält. In dieser Situation sollten Sie die Mindestwerte verringern, um so aussagekräftigere Ergebnisse zu erzielen.

CHAID-Kriterien

Abbildung 1-6
Dialogfeld "Kriterien", Registerkarte "CHAID"

Bei den Methoden CHAID und Exhaustive CHAID können Sie Folgendes steuern:

Signifikanzniveau. Legen Sie den Signifikanzwert für das Aufteilen von Knoten und das Zusammenführen von Kategorien fest. Bei beiden Kriterien liegt das Standard-Signifikanzniveau bei 0,05.

- Beim Aufteilen von Knoten muss der Wert größer als 0 und kleiner als 1 sein. Bei niedrigeren Werten entstehen Bäume mit weniger Knoten.
- Beim Zusammenführen von Kategorien muss der Wert größer als 0 und kleiner oder gleich 1 sein. Wenn ein Zusammenführen der Kategorien unterbunden werden soll, legen Sie den Wert 1 fest. Bei einer metrischen unabhängigen Variablen bedeutet dies, dass die Anzahl der Kategorien für die Variable im fertigen Baum der angegebenen Anzahl an Intervallen entspricht (Standardwert: 10). Für weitere Informationen siehe Metrische Intervalle für die CHAID-Analyse auf S. 11.

Chi-Quadrat-Statistik. Bei ordinalen abhängigen Variablen wird der Chi-Quadrat-Wert, mit dem das Aufteilen von Knoten und das Zusammenführen von Kategorien bestimmt wird, mithilfe der Likelihood-Quotienten-Methode berechnet. Bei nominalen abhängigen Variablen können Sie die Methode auswählen:

- **Pearson.** Diese Methode liefert schnellere Berechnungen, sollte bei kleineren Stichproben jedoch nur nach sorgfältiger Erwägung verwendet werden. Dies ist die Standardmethode.
- **Likelihood-Quotient.** Diese Methode ist stabiler als die Pearson-Methode; die Berechnungen nehmen jedoch mehr Zeit in Anspruch. Diese Methode eignet sich ideal für kleine Stichproben.

Modellschätzung. Bei nominalen und ordinalen abhängigen Variablen können Sie Folgendes festlegen:

- **Die maximale Anzahl von Iterationsschritten.** Der Standardwert ist 100. Wenn der Baum nicht mehr weiter aufgebaut wird, weil die maximale Anzahl an Iterationen erreicht ist, können Sie den Maximalwert erhöhen oder auch ein oder mehrere Kriterien ändern, die den Aufbau des Baums steuern.
- **Mindeständerung bei den erwarteten Zellhäufigkeiten.** Der Wert muss größer als 0 und kleiner als 1 sein. Der Standardwert ist 0,05. Bei niedrigeren Werten entstehen Bäume mit weniger Knoten.

Signifikanzwerte mit der Bonferroni-Methode anpassen. Bei Mehrfachvergleichen werden die Signifikanzwerte für die Zusammenführungs- und Aufteilungskriterien mithilfe der Bonferroni-Methode angepasst. Dies ist die Standardeinstellung.

Erneute Aufteilung zusammengeführter Kategorien innerhalb eines Knotens zulassen. Sofern Sie das Zusammenführen von Kategorien nicht explizit unterbinden, werden Kategorien mit unabhängigen Variablen (Einflussvariablen) nach Möglichkeit zusammengeführt, um so den einfachsten Baum zu bilden, der das Modell beschreibt. Bei dieser Option können zusammengeführte Kategorien eigenständig durch die Prozedur erneut aufgeteilt werden, wenn hierdurch eine bessere Lösung entstünde.

Metrische Intervalle für die CHAID-Analyse

Abbildung 1-7
Dialogfeld "Kriterien," Registerkarte "Intervalle"

Bei der CHAID-Analyse werden metrische unabhängige Variablen (Einflussvariablen) vor der Analyse stets in diskrete Gruppen eingeteilt (z. B. 0–10, 11–20, 21–30 usw.). Sie können die anfängliche und maximale Anzahl der Gruppen steuern (unter Umständen werden aufeinander folgende Gruppen nach der ursprünglichen Aufteilung jedoch wieder zusammengeführt):

- **Feste Zahl.** Alle metrischen unabhängigen Variablen werden zunächst in dieselbe Anzahl an Gruppen eingeteilt. Der Standardwert ist 10.
- **Benutzerdefiniert.** Jede metrische unabhängige Variable wird zunächst in die Anzahl der Gruppen eingeteilt, die für die betreffende Variable angegeben sind.

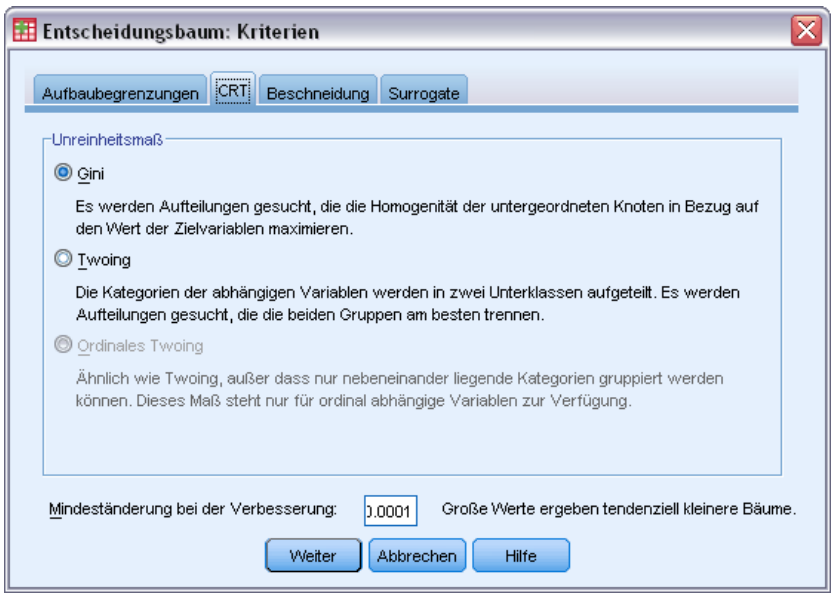
So legen Sie die Intervalle für metrische unabhängige Variablen fest:

- ▶ Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" mindestens eine metrische unabhängige Variable aus.
- ▶ Wählen Sie als Aufbaumethode die Option CHAID oder Exhaustive CHAID.
- ▶ Klicken Sie auf Kriterien.
- ▶ Klicken Sie auf die Registerkarte Intervalle.

Bei der CRT- und QUEST-Analyse werden nur binäre Aufteilungen verwendet und die metrischen und ordinalen unabhängigen Variablen werden auf dieselbe Weise behandelt. Es ist also nicht möglich, eine Intervallanzahl für die metrischen unabhängigen Variablen festzulegen.

CRT-Kriterien

Abbildung 1-8
Dialogfeld "Kriterien", Registerkarte "CRT"


Entscheidungsbaum: Kriterien
Aufbaubegrenzungen
CRT
Beschneidung
Surrogate
Unreinheitsmaß
Gini
Es werden Aufteilungen gesucht, die die Homogenität der untergeordneten Knoten in Bezug auf den Wert der Zielvariablen maximieren.
Twoing
Die Kategorien der abhängigen Variablen werden in zwei Unterklassen aufgeteilt. Es werden Aufteilungen gesucht, die die beiden Gruppen am besten trennen.
Ordinales Twoing
Ähnlich wie Twoing, außer dass nur nebeneinander liegende Kategorien gruppiert werden können. Dieses Maß steht nur für ordinal abhängige Variablen zur Verfügung.
Mindeständerung bei der Verbesserung:
0.0001
Große Werte ergeben tendenziell kleinere Bäume.
Weiter
Abbrechen
Hilfe

Bei der CRT-Aufbaumethode wird die Homogenität innerhalb der Knoten angestrebt. Das Ausmaß, in dem ein Knoten von einer homogenen Untergruppe von Fällen abweicht, ist ein Hinweis auf **Unreinheit**. Beispiel: Ein Endknoten, in dem alle Fälle denselben Wert für die

abhängige Variable aufweisen, ist ein homogener Knoten. Eine weitere Aufteilung ist nicht nötig, weil der Knoten bereits “rein” ist.

Sie können die Methode zum Messen der Unreinheit bestimmen und auch den Rückgang in der Unreinheit angeben, der mindestens erreicht werden muss, damit die Knoten aufgeteilt werden.

Unreinheitsmaß. Bei metrischen abhängigen Variablen wird das LSD-Unreinheitsmaß (Least-Squared Deviation, kleinste quadratische Abweichung) verwendet. Dieser Wert wird als Varianz innerhalb der Knoten berechnet und ggf. gemäß der Häufigkeitsgewichtungen oder der Einflusswerte angepasst.

Bei kategorialen (nominalen, ordinalen) abhängigen Variablen stehen die folgenden Unreinheitsmaße zur Auswahl:

- **Gini.** Die Aufteilungen maximieren die Homogenität der untergeordneten Knoten im Hinblick auf den Wert der abhängigen Variable. Das Gini-Maß beruht auf den quadratischen Wahrscheinlichkeiten für die Zugehörigkeit zu einer Kategorie der abhängigen Variable. Der Mindestwert (Null) wird erreicht, sobald alle Fälle in einem Knoten in eine einzige Kategorie fallen. Dies ist das Standardmaß.
- **Twoing.** Die Kategorien der abhängigen Variablen werden in zwei Unterklassen gruppiert. Die Aufteilungen bewirken die bestmögliche Trennung der beiden Gruppen.
- **Ordinales Twoing.** Dieses Maß entspricht weitgehend dem Twoing, mit der Ausnahme, dass nur nebeneinander liegende Kategorien gruppiert werden können. Dieses Maß steht nur bei ordinalen abhängigen Variablen zur Verfügung.

Mindeständerung bei der Verbesserung. Dies ist der mindestens erforderliche Rückgang der Unreinheit für das Aufteilen eines Knotens. Der Standardwert ist 0,0001. Bei höheren Werten entstehen Bäume mit weniger Knoten.

QUEST-Kriterien

Abbildung 1-9
Dialogfeld "Kriterien", Registerkarte "QUEST"

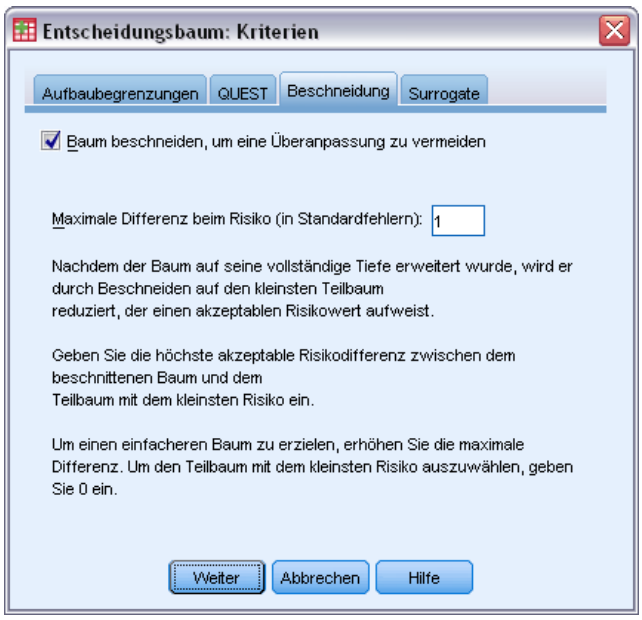
Bei der QUEST-Methode können Sie das Signifikanzniveau für das Aufteilen von Knoten festlegen. Die Knoten können nur dann mit einer unabhängigen Variablen aufgeteilt werden, wenn das Signifikanzniveau kleiner oder gleich dem angegebenen Wert ist. Der Wert muss größer als 0 und kleiner als 1 sein. Der Standardwert ist 0,05. Bei kleineren Werten werden mehr unabhängige Variablen aus dem endgültigen Modell ausgeschlossen.

So legen Sie die QUEST-Kriterien fest:

- Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine nominale abhängige Variable aus.
- Wählen Sie als Aufbaumethode die Option QUEST.
- Klicken Sie auf Kriterien.
- Klicken Sie auf die Registerkarte QUEST.

Beschneiden von Bäumen

Abbildung 1-10
Dialogfeld "Kriterien", Registerkarte "Beschneidung"


Entscheidungsbaum: Kriterien
Aufbaubegrenzungen
QUEST
Beschneidung
Surrogate
Baum beschneiden, um eine Überanpassung zu vermeiden
Maximale Differenz beim Risiko (in Standardfehlern):
1
Nachdem der Baum auf seine vollständige Tiefe erweitert wurde, wird er durch Beschneiden auf den kleinsten Teilbaum reduziert, der einen akzeptablen Risikowert aufweist.
Geben Sie die höchste akzeptable Risikodifferenz zwischen dem beschnittenen Baum und dem Teilbaum mit dem kleinsten Risiko ein.
Um einen einfacheren Baum zu erzielen, erhöhen Sie die maximale Differenz. Um den Teilbaum mit dem kleinsten Risiko auszuwählen, geben Sie 0 ein.
Weiter
Abbrechen
Hilfe

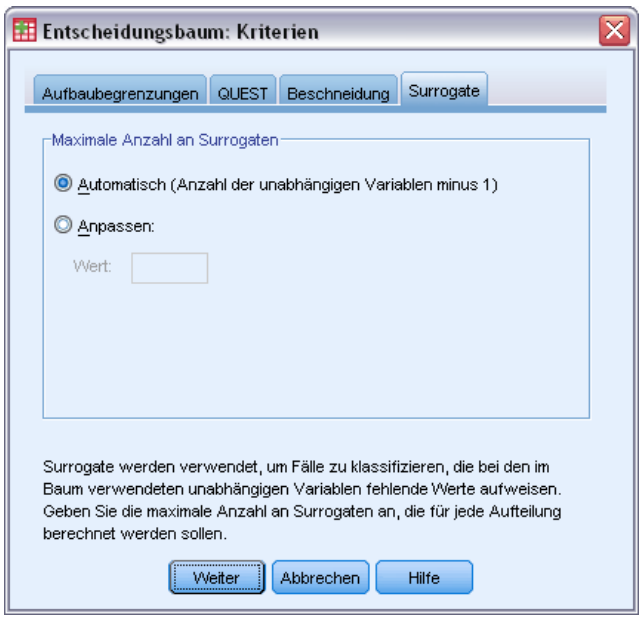
Bei der CRT- und der QUEST-Methode können Sie ein Überfüllen des Modells vermeiden, indem Sie den Baum **zuschneiden**: Der Baum wächst, bis die Kriterien für das Anhalten erfüllt sind. Anschließend wird der Baum automatisch gemäß der angegebenen maximalen Risikodifferenz auf den kleinsten Teilbaum beschnitten. Der Risikowert wird in Standardfehlern ausgedrückt. Der Standardwert ist 1. Der Wert muss positiv oder gleich Null sein. Um den Teilbaum mit dem geringstmöglichen Risiko zu erzielen, geben Sie den Wert 0 an.

Beschneiden im Vergleich mit dem Ausblenden von Knoten

Bei einem beschnittenen Baum sind alle Knoten, die aus dem Baum herausgeschnitten wurden, im endgültigen Baum nicht mehr verfügbar. Sie können zwar ausgewählte untergeordnete Knoten im fertigen Baum interaktiv ein- und ausblenden; es ist jedoch nicht möglich, Knoten anzeigen zu lassen, die beim Erstellen des Baums beschnitten wurden. Für weitere Informationen siehe Baumeditor in Kapitel 2 auf S. 41.

Surrogate

Abbildung 1-11
Dialogfeld "Kriterien", Registerkarte "Surrogate"


Entscheidungsbaum: Kriterien
Aufbaubegrenzungen
QUEST
Beschneidung
Surrogate
Maximale Anzahl an Surrogaten
Automatisch (Anzahl der unabhängigen Variablen minus 1)
Anpassen:
Wert:
Surrogate werden verwendet, um Fälle zu klassifizieren, die bei den im Baum verwendeten unabhängigen Variablen fehlende Werte aufweisen. Geben Sie die maximale Anzahl an Surrogaten an, die für jede Aufteilung berechnet werden sollen.
Weiter
Abbrechen
Hilfe

Bei CRT und QUEST können **Surrogate** für unabhängige Variablen (Einflussvariablen) verwendet werden. In Situationen, in denen der Wert für die betreffende Variable fehlt, werden andere unabhängige Variablen, die einen hohen Grad an Zusammenhang mit der ursprünglichen Variable besitzen, zur Klassifizierung herangezogen. Diese alternativen Einflussvariablen werden als Surrogate bezeichnet. Sie können die maximal zulässige Anzahl an Surrogaten für das Modell festlegen.

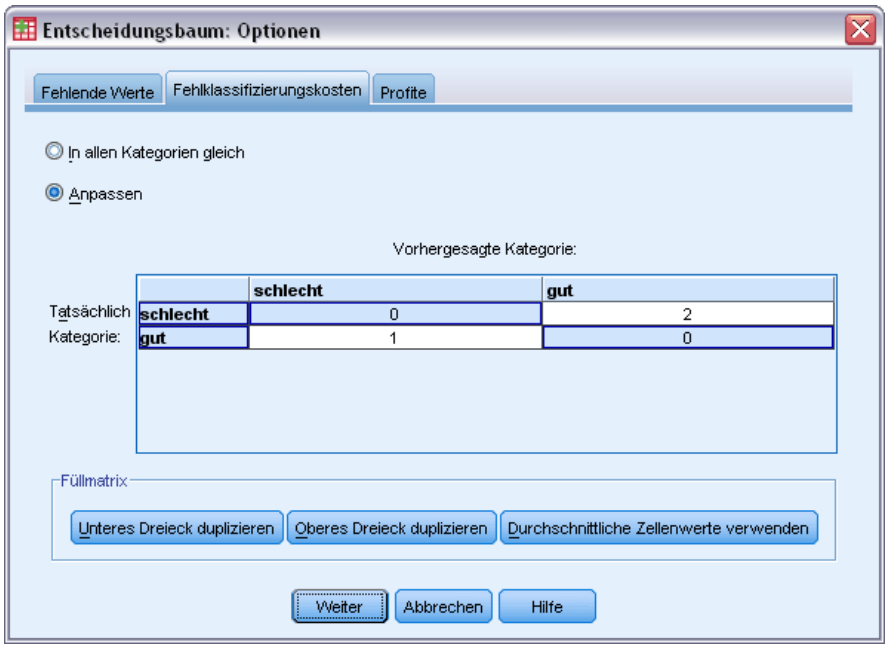
- Standardmäßig ist die maximale Anzahl an Surrogaten um 1 kleiner als die Anzahl der unabhängigen Variablen. Für eine unabhängige Variable kann also jede andere unabhängige Variable als Surrogat verwendet werden.
- Sollen keine Surrogate im Modell verwendet werden, geben Sie den Wert 0 als Anzahl der Surrogate an.

Optionen

Die tatsächlich verfügbaren Optionen sind abhängig von der Aufbaumethode, dem Messniveau der abhängigen Variablen und/oder dem Vorhandensein definierter Wertelabel für die Werte der abhängigen Variable.

Fehlklassifizierungskosten

Abbildung 1-12
Dialogfeld "Optionen", Registerkarte "Fehlklassifizierungskosten"



Bei kategorialen (nominalen, ordinalen) abhängigen Variablen können Sie mit den Fehlklassifizierungskosten die relative Strafe für die fehlerhafte Klassifizierung angeben. Beispiel:

- Die Kosten, wenn einem kreditwürdigen Kunden ein Darlehen verweigert wird, unterscheiden sich in der Regel von den Kosten, wenn ein Kunde ein Darlehen erhält und dann damit in Verzug gerät.
- Die Kosten für die Fehlklassifizierung einer Person mit einem hohen Risiko für Herzerkrankungen als Person mit niedrigem Risiko sind wahrscheinlich deutlich höher, als wenn eine Person mit niedrigem Risiko fälschlicherweise mit einem hohen Risiko klassifiziert würde.
- Die Kosten für den Versand einer Werbesendung an eine Person, die wahrscheinlich nicht reagieren wird, sind relativ gering; die Kosten, wenn die Werbesendung nicht an eine Person geht, die wahrscheinlich reagiert hätte, sind dagegen deutlich höher (was den entgangenen Umsatz angeht).

"Fehlklassifizierungskosten" und Wertelabels

Dieses Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen ein Wertelabel besitzen.

So legen Sie die Fehlklassifizierungskosten fest:

▶ Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertelabels aus.

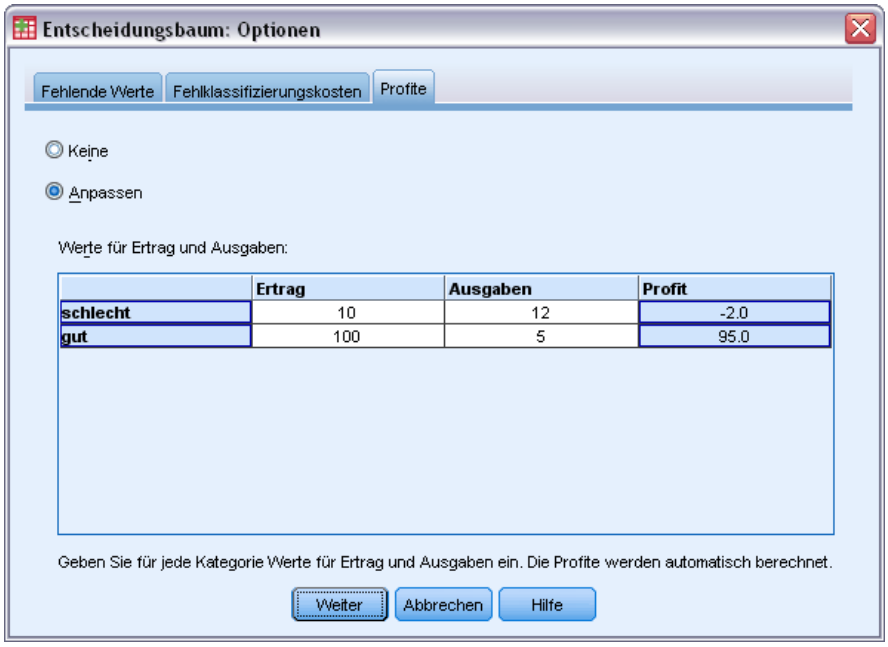
- ▶ Klicken Sie auf Optionen.
- ▶ Klicken Sie auf die Registerkarte Fehlklassifizierungskosten.
- ▶ Klicken Sie auf Benutzerdefiniert.
- ▶ Geben Sie mindestens einen Wert für die Fehlklassifizierungskosten in das Gitter ein. Die Werte müssen positiv oder gleich Null sein. (Richtige Klassifizierungen, auf der Diagonalen dargestellt, sind stets gleich 0.)

Füllmatrix. Häufig sollen die Kosten symmetrisch sein: Die Kosten für die Fehlklassifizierung von A als B sind genauso hoch wie die Kosten für die Fehlklassifizierung von B als A. Die folgenden Steuerungen erleichtern das Anlegen einer symmetrischen Kostenmatrix:

- **Unteres Dreieck duplizieren.** Kopiert Werte aus dem unteren Dreieck der Matrix (unterhalb der Diagonalen) in die entsprechenden Zellen oberhalb des Dreiecks.
- **Oberes Dreieck duplizieren.** Kopiert Werte aus dem oberen Dreieck der Matrix (oberhalb der Diagonalen) in die entsprechenden Zellen unterhalb des Dreiecks.
- **Durchschnittliche Zellenwerte verwenden.** Für jede Zelle in beiden Hälften der Matrix wird der Durchschnitt aus den beiden Werten (im oberen und unteren Dreieck) gebildet und anstelle der ursprünglichen beiden Werte eingesetzt. Beispiel: Die Fehlklassifizierung von A als B verursacht Kosten in Höhe von 1 und die Kosten für die Fehlklassifizierung von B als A betragen 3. Beide Werte werden somit durch den Durchschnitt (1+3)/2 = 2 ersetzt.

Profite

Abbildung 1-13
Dialogfeld "Optionen", Registerkarte "Profite"



Bei kategorialen abhängigen Variablen können Sie den verschiedenen Ebenen jeweils Werte für Verkaufserlöse und Aufwendungen zuweisen.

- Der Profit ergibt sich aus der Berechnung Verkaufserlöse minus Aufwendungen.
- Die Profitwerte beeinflussen die Werte für den durchschnittlichen Profit und den Anlageertrag (ROI) in den Gewinntabellen. Die grundlegende Baummodellstruktur bleibt unverändert.
- Die Werte für Verkaufserlöse und Aufwendungen müssen numerisch sein und müssen für alle im Gitter angezeigten Kategorien der abhängigen Variablen festgelegt werden.

"Profite" und Wertelabels

In diesem Dialogfeld sind definierte Wertelabels für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen ein Wertelabel besitzen.

So geben Sie die Gewinne an:

▶ Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertelabels aus.

▶ Klicken Sie auf Optionen.

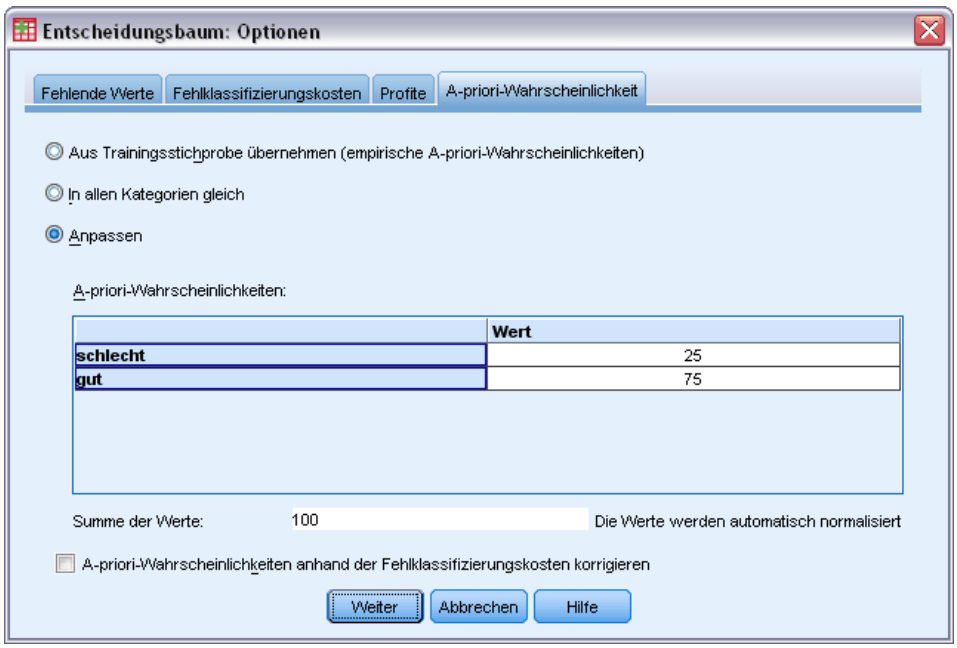
▶ Klicken Sie auf die Registerkarte Profite.

▶ Klicken Sie auf Benutzerdefiniert.

▶ Geben Sie die Werte für Verkaufserlöse und Aufwendungen für alle im Gitter aufgeführten Kategorien der abhängigen Variablen ein.

A-priori-Wahrscheinlichkeit

Abbildung 1-14
Dialogfeld "Optionen", Registerkarte "A-priori-Wahrscheinlichkeiten"



Bei CRT- und QUEST-Bäumen mit kategorialen abhängigen Variablen können Sie A-priori-Wahrscheinlichkeiten für die Gruppenzugehörigkeit angeben. **A-priori-Wahrscheinlichkeiten** sind eine Schätzung der gesamten relativen Häufigkeit für jede Kategorie der abhängigen Variable, die aufgestellt wird, noch bevor die Werte der unabhängigen Variablen (Einflussvariablen) bekannt sind. Mithilfe von A-priori-Wahrscheinlichkeiten können Sie den Aufbau des Baums durch Daten in der Stichprobe korrigieren, die nicht repräsentativ für die Gesamtheit als Ganzes sind.

Aus Trainingsstichprobe übernehmen (empirische A-priori-Wahrscheinlichkeiten). Aktivieren Sie diese Einstellung, wenn die Verteilung der Variablenwerte in der Datendatei repräsentativ für die Verteilung in der Gesamtheit ist. Bei der Split-Sample-Validierung wird die Verteilung der Fälle in der Trainingsstichprobe herangezogen.

Hinweis: Bei der Split-Sample-Validierung werden die Fälle nach dem Zufallsprinzip in die Trainingsstichprobe aufgenommen. Die eigentliche Verteilung der Fälle in der Trainingsstichprobe ist daher im Voraus nicht bekannt. Für weitere Informationen siehe Validierung auf S. 7.

In allen Kategorien gleich. Aktivieren Sie diese Einstellung, wenn die Kategorien der abhängigen Variablen in der Gesamtheit gleichmäßig repräsentiert sind. Beispiel: Es liegen vier Kategorien vor und auf jede Kategorie entfallen etwa 25 % der Fälle.

Benutzerdefiniert. Geben Sie je einen positiven Wert (oder den Wert 0) für jede im Gitter aufgeführte Kategorie der abhängigen Variablen ein. Die Werte können Anteile, Prozentsätze oder Häufigkeitszählungen umfassen oder auch andere Werte, die die Verteilung der Werte in den Kategorien wiedergeben.

A-priori-Wahrscheinlichkeiten anhand der Fehlklassifizierungskosten korrigieren. Wenn Sie benutzerdefinierte Fehlklassifizierungskosten definieren, können Sie die A-priori-Wahrscheinlichkeiten anhand dieser Kosten anpassen. Für weitere Informationen siehe Fehlklassifizierungskosten auf S. 17.

"Profite" und Wertelabels

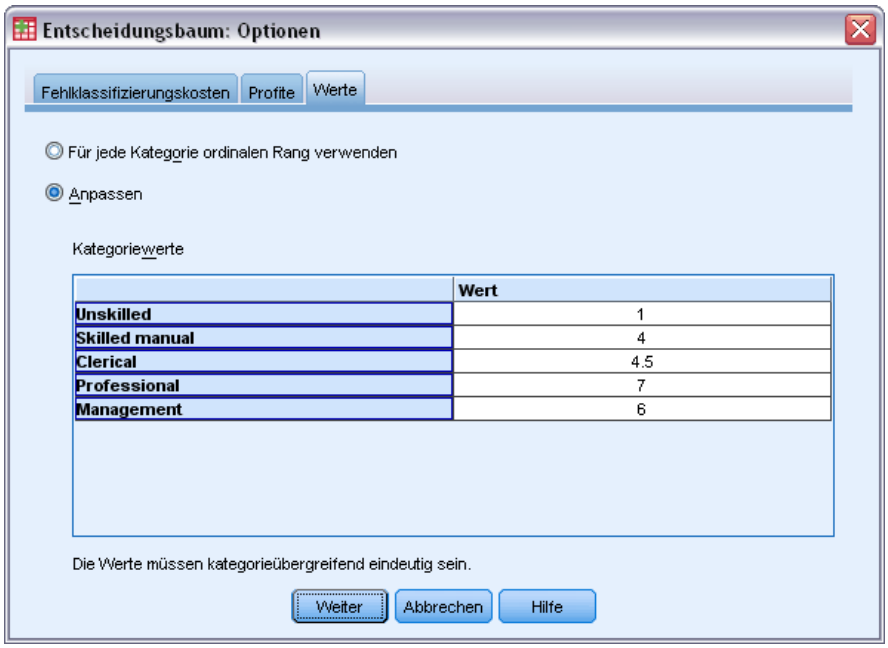
In diesem Dialogfeld sind definierte Wertelabels für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen ein Wertelabel besitzen.

So legen Sie A-priori-Wahrscheinlichkeiten fest:

- Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertelabels aus.
- Wählen Sie als Aufbaumethode die Option CRT oder QUEST.
- Klicken Sie auf Optionen.
- Klicken Sie auf die Registerkarte A-priori-Wahrscheinlichkeiten.

Werte

Abbildung 1-15
Dialogfeld "Optionen", Registerkarte "Werte"



Bei CHAID und Exhaustive CHAID mit einer ordinalen abhängigen Variablen können Sie benutzerdefinierte Score-Werte für die einzelnen Kategorien der abhängigen Werte zuweisen. Die Score-Werte definieren die Reihenfolge für die Kategorien der abhängigen Variablen und die

Distanz zwischen diesen Kategorien. Mithilfe der Score-Werte können Sie die relative Distanz zwischen ordinalen Werten vergrößern oder verkleinern sowie die Reihenfolge der Werte ändern.

- **Für jede Kategorie ordinalen Rang verwenden.** Die niedrigste Kategorie der abhängigen Variablen erhält den Score-Wert 1, die nächsthöhere Kategorie den Score-Wert 2 usw. Dies ist die Standardeinstellung.
- **Benutzerdefiniert.** Geben Sie je einen numerischen Score-Wert für jede im Gitter aufgeführte Kategorie der abhängigen Variablen ein.

Beispiel

Wertbeschriftung	Originalwert	Wert
Ungelernt	1	1
Gelernt/Werkstatt	2	4
Verwaltung	3	4.5
Professional	4	7
Management	5	6

- Die Score-Werte vergrößern die relative Distanz zwischen *Ungelernt* und *Gelernt/Werkstatt* und verringern die relative Distanz zwischen *Gelernt/Werkstatt* und *Verwaltung*.
- Die Score-Werte kehren die Reihenfolge von *Management* und *Fachkraft* um.

"Werte" und Wertlabels

In diesem Dialogfeld sind definierte Wertelabels für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen ein Wertelabel besitzen.

So legen Sie Score-Werte fest:

▶ Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine ordinale abhängige Variable mit mindestens zwei definierten Wertelabels aus.

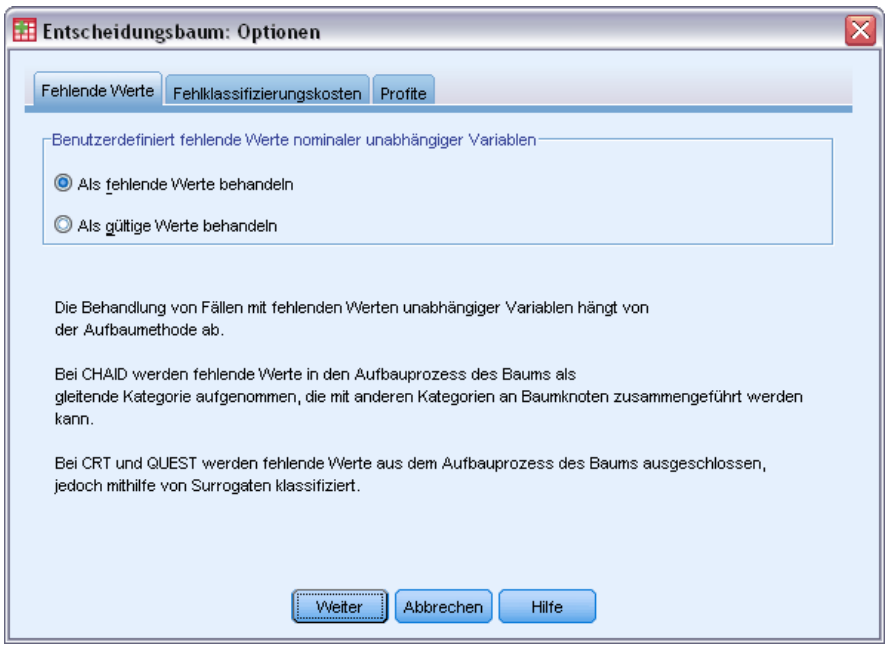
▶ Wählen Sie als Aufbaumethode die Option CHAID oder Exhaustive CHAID.

▶ Klicken Sie auf Optionen.

▶ Klicken Sie auf die Registerkarte Werte.

Fehlende Werte definieren

Abbildung 1-16
Dialogfeld "Optionen", Registerkarte "Fehlende Werte"


Entscheidungsbaum: Optionen
Fehlende Werte
Fehlklassifizierungskosten
Profite
Benutzerdefiniert fehlende Werte nominaler unabhängiger Variablen
Als fehlende Werte behandeln
Als gültige Werte behandeln
Die Behandlung von Fällen mit fehlenden Werten unabhängiger Variablen hängt von der Aufbaumethode ab.
Bei CHAID werden fehlende Werte in den Aufbauprozess des Baums als gleitende Kategorie aufgenommen, die mit anderen Kategorien an Baumknoten zusammengeführt werden kann.
Bei CRT und QUEST werden fehlende Werte aus dem Aufbauprozess des Baums ausgeschlossen, jedoch mithilfe von Surrogaten klassifiziert.
Weiter
Abbrechen
Hilfe

Auf der Registerkarte "Fehlende Werte" steuern Sie die Behandlung benutzerdefiniert fehlender Werte für nominale unabhängige Variablen (Einflussvariablen).

- Benutzerdefiniert fehlende Werte für ordinale und metrische Variablen werden bei den verschiedenen Aufbaumethoden auf unterschiedliche Weise behandelt.
- Die Behandlung nominaler abhängiger Variablen wird im Dialogfeld "Kategorien" festgelegt. Für weitere Informationen siehe Auswählen von Kategorien auf S. 6.
- Bei ordinalen und metrischen abhängigen Variablen werden Fälle, bei denen systemdefiniert oder benutzerdefiniert fehlende Werte vorliegen, stets ausgeschlossen.

Als fehlende Werte behandeln. Benutzerdefiniert fehlende Werte werden wie systemdefiniert fehlende Werte behandelt. Systemdefiniert fehlende Werte werden bei den verschiedenen Aufbaumethoden auf unterschiedliche Weise behandelt.

Als gültige Werte behandeln. Benutzerdefinert fehlende Werte bei nominalen unabhängigen Variablen werden beim Aufbau und bei der Klassifizierung des Baums als normale Werte behandelt.

Methodenspezifische Regeln

Einige (jedoch nicht alle) Werte für eine unabhängige Variable fehlen system- oder benutzerdefiniert:

- Bei CHAID und Exhaustive CHAID werden system- und benutzerdefiniert fehlende Werte für eine unabhängige Variable als eine einzige, kombinierte Kategorie in die Analyse aufgenommen. Bei metrischen und ordinalen unabhängigen Variablen werden mit den Algorithmen zunächst Kategorien mithilfe gültiger Werte erzeugt. Anschließend wird entschieden, ob die fehlende Kategorie mit der ähnlichsten (gültigen) Kategorie zusammengeführt oder als separate Kategorie beibehalten werden soll.
- Bei CRT und QUEST werden Fälle, bei denen Werte für eine unabhängige Variable fehlen, aus dem Vorgang des Baumaufbaus ausgeschlossen. Falls Surrogate in der Methode eingeschlossen sind, werden diese Fälle allerdings mithilfe von Surrogaten klassifiziert. Für nominale benutzerdefiniert fehlende Werte, die als fehlend behandelt werden, gilt dieselbe Vorgehensweise. Für weitere Informationen siehe Surrogate auf S. 16.

So bestimmen Sie die Behandlung für nominale, unabhängige, benutzerdefiniert fehlende Werte:

- ▶ Wählen Sie im Hauptdialogfeld “Entscheidungsbaum” mindestens eine nominale unabhängige Variable aus.
- ▶ Klicken Sie auf Optionen.
- ▶ Klicken Sie auf die Registerkarte Fehlende Werte.

Speichern der Modelldaten

Abbildung 1-17
Dialogfeld “Speichern”

Sie können die Daten aus dem Modell als Variablen in der Arbeitsdatei ablegen und auch das gesamte Modell im XML-Format (PMML) in eine externe Datei speichern.

Gespeicherte Variablen

Endknotennummer. Endknoten, dem die einzelnen Fälle zugewiesen sind. Der Wert ist die Baumknotennummer.

Vorhergesagter Wert. Klasse (Gruppe) oder Wert für die abhängige Variable, der durch das Modell vorhergesagt wurde.

Vorhergesagte Wahrscheinlichkeiten. Wahrscheinlichkeit, die mit der Vorhersage des Modells verbunden ist. Für jede Kategorie der abhängigen Variablen wird je eine Variable gespeichert. Nicht verfügbar für metrische abhängige Variablen.

Stichprobenzuweisungen (Training/Tests). Diese Variable zeigt bei der Split-Sample-Validierung, ob ein Fall in der Trainings- oder in der Teststichprobe verwendet wurde. Bei der Trainingsstichprobe ist der Wert gleich 1, bei der Teststichprobe dagegen gleich 0. Nur verfügbar, wenn die Split-Sample-Validierung ausgewählt ist. Für weitere Informationen siehe Validierung auf S. 7.

Baummodell als XML exportieren

Sie können das gesamte Baummodell im XML-Format (PMML) speichern. *SmartScore* und PASW StatisticsServer (gesondertes Produkt) können anhand dieser Modelldatei die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden.

Trainingsstichprobe. Schreibt das Modell in die angegebene Datei. Bei Bäumen mit Split-Sample-Validierung ist dies das Modell für die Trainingsstichprobe.

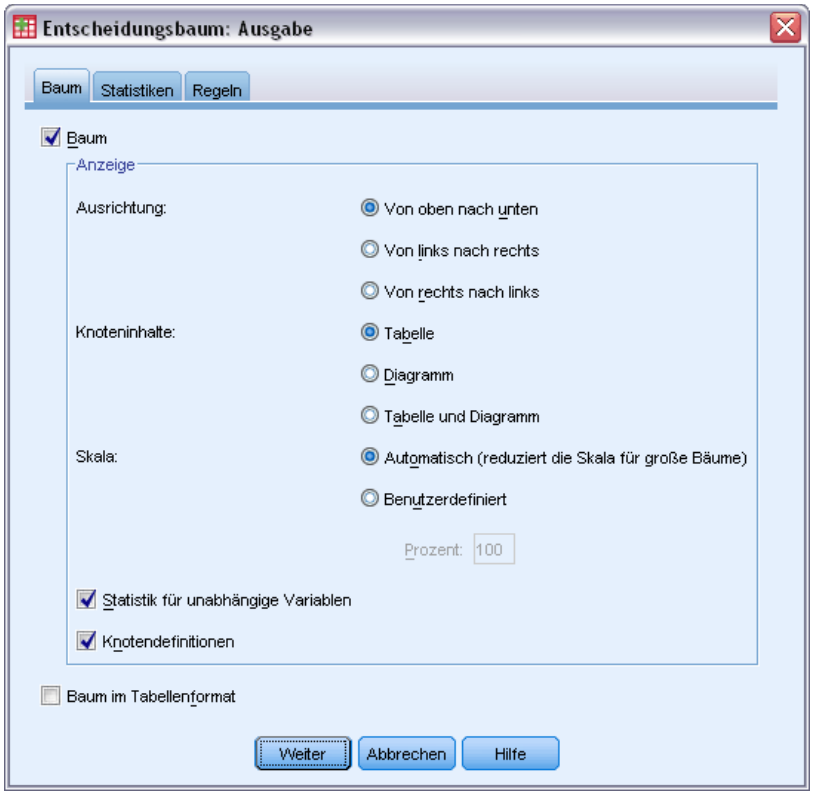
Teststichprobe. Schreibt das Modell für die Teststichprobe in die angegebene Datei. Nur verfügbar, wenn die Split-Sample-Validierung ausgewählt ist.

Ausgabe

Die verfügbaren Ausgabeoptionen sind abhängig von der Aufbaumethode, dem Messniveau der abhängigen Variablen und anderen Einstellungen.

Baumanzeige

Abbildung 1-18
Dialogfeld "Ausgabe", Registerkarte "Baum"



Sie können das anfängliche Erscheinungsbild des Baums steuern oder auch die Baumanzeige ganz unterdrücken.

Baum. Standardmäßig wird das Baumdiagramm in der Ausgabe im Viewer dargestellt. Soll das Baumdiagramm nicht in der Ausgabe angezeigt werden, deaktivieren Sie diese Option.

Anzeigen. Diese Optionen steuern das anfängliche Erscheinungsbild des Baumdiagramms im Viewer. Diese Attribute können außerdem geändert werden, indem Sie den erzeugten Baum bearbeiten.

- **Ausrichtung.** Der Baum kann wahlweise auf dem Kopf stehend (mit dem Stammknoten an oberster Stelle), von links nach rechts oder von rechts nach links angezeigt werden.
- **Knoteninhalte.** Die Knoten können Tabellen und/oder Diagramme enthalten. Bei kategorialen abhängigen Variablen zeigen die Tabellen die Häufigkeitszählungen und die Prozentsätze; die Diagramme bestehen dabei aus Balkendiagrammen. Bei metrischen abhängigen Variablen zeigen die Tabellen die Mittelwerte, die Standardabweichungen, die Anzahl der Fälle und die vorhergesagten Werte. Die Diagramme bestehen dabei aus Histogrammen.
- **Skala.** Standardmäßig werden große Bäume so skaliert, dass der gesamte Baum auf der Seite dargestellt werden kann. Sie können eine benutzerdefinierte Skalierung bis 200 % angeben.

- **Statistik für unabhängige Variablen.** Bei CHAID und Exhaustive CHAID umfassen die Statistiken den *F*-Wert (metrische abhängige Variablen) bzw. den Chi-Quadrat-Wert (kategoriale abhängige Variablen), außerdem den Signifikanzwert und die Freiheitsgrade. Bei CRT wird der Verbesserungswert angezeigt. Bei QUEST werden der *F*-Wert, der Signifikanzwert und die Freiheitsgrade (für metrische und ordinale unabhängige Variablen) bzw. der Chi-Quadrat-Wert, der Signifikanzwert und die Freiheitsgrade (für nominale unabhängige Variablen) angezeigt.
- **Knotendefinitionen.** Die Knotendefinitionen zeigen den Wert oder die Werte der unabhängigen Variablen bei jeder Knotenaufteilung.

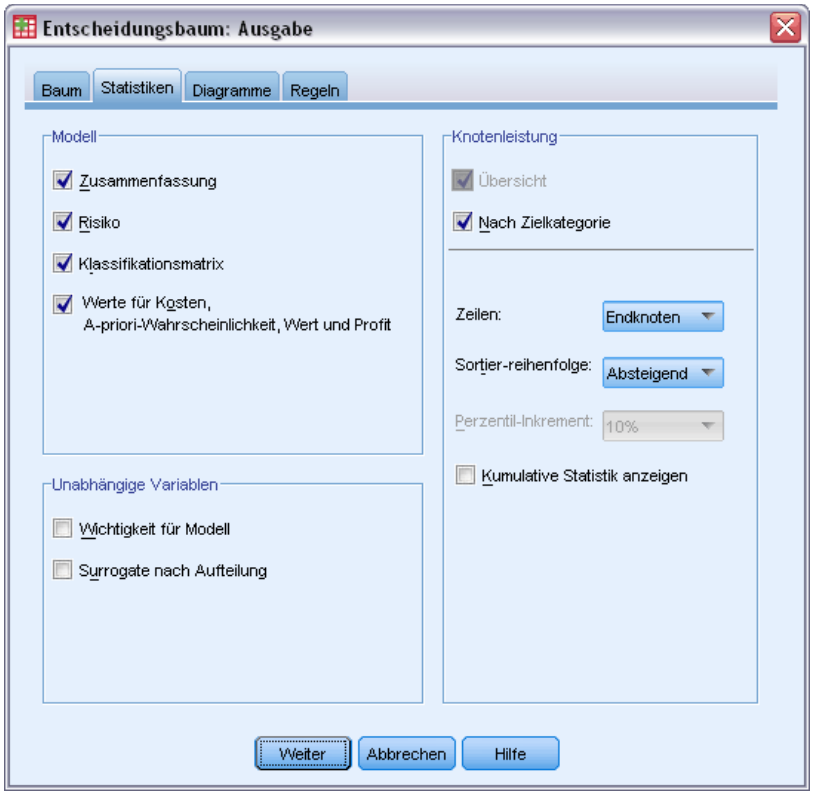
Baum im Tabellenformat. Zusammenfassende Angaben für jeden Knoten im Baum: Nummer des übergeordneten Knotens, Statistik für unabhängige Variablen, Wert(e) der unabhängigen Variablen für den Knoten, Mittelwert und Standardabweichung für metrische abhängige Variablen bzw. Zählungen und Prozentsätze für kategoriale abhängige Variablen.

Abbildung 1-19
Baum im Tabellenformat

	schlecht		gut		Gesamt				Primäre unabhängige Variable				
Knoten	N	Prozent	N	Prozent	N	Prozent	Vorhergesagte Kategorie	Übergeordneter Knoten	Variable	Sig.[a]	Chi-Quadrat	df	Aufteilungwerte
0	1020	41,4%	1444	58,6%	2464	100,0%	gut						
1	454	82,1%	99	17,9%	553	22,4%	schlecht	0	Einkommen in Kategorien	,000	662,457	2	<= niedrig
2	476	42,0%	658	58,0%	1134	46,0%	gut	0	Einkommen in Kategorien	,000	662,457	2	(niedrig, mittel]
3	90	11,6%	687	88,4%	777	31,5%	gut	0	Einkommen in Kategorien	,000	662,457	2	> mittel
4	422	56,7%	322	43,3%	744	30,2%	schlecht	2	Anzahl an Kreditkarten	,000	193,113	1	5 oder mehr
5	54	13,8%	336	86,2%	390	15,8%	gut	2	Anzahl an Kreditkarten	,000	193,113	1	weniger als 5
6	80	17,6%	375	82,4%	455	18,5%	gut	3	Anzahl an Kreditkarten	,000	38,587	1	5 oder mehr
7	10	3,1%	312	96,9%	322	13,1%	gut	3	Anzahl an Kreditkarten	,000	38,587	1	weniger als 5
8	211	80,8%	50	19,2%	261	10,6%	schlecht	4	Alter	,000	95,299	1	<= 28,079205818990676
9	211	43,7%	272	56,3%	483	19,6%	gut	4	Alter	,000	95,299	1	> 28,079205818990676

Statistik

Abbildung 1-20
Dialogfeld "Optionen", Registerkarte "Statistik"



Die verfügbaren Statistiktabellen sind abhängig vom Messniveau der abhängigen Variable, von der Aufbaumethode und anderen Einstellungen.

Modell

Zusammenfassung. Die Zusammenfassung zeigt die verwendete Methode, die Variablen, die im Modell berücksichtigt sind, sowie die Variablen, die zwar angegeben, jedoch nicht in das Modell aufgenommen wurden.

Abbildung 1-21
Modellzusammenfassungstabelle

Spezifikationen	Aufbaumethode	CHAID
	Abhängige Variable	Kreditrating
	Unabhängige Variablen	Alter, Einkommen in Kategorien, Anzahl an Kreditkarten, Ausbildung, Autodarlehen
	Validierung	NONE
	Maximale Baumtiefe	3
	Mindestanzahl der Fälle im übergeordneten Knoten	400
	Mindestanzahl der Fälle im untergeordneten Knoten	200
Ergebnisse	Aufgenommene unabhängige Variablen	Einkommen in Kategorien, Anzahl an Kreditkarten, Alter
	Anzahl der Knoten	10
	Anzahl der Endknoten	6
	Tiefe	3

Risiko. Risikoschätzung und zugehöriger Standardfehler. Maß für die Vorhersagegenauigkeit des Baums.

- Bei kategorialen abhängigen Variablen ist die Risikoschätzung der Anteil der Fälle, die nach der Anpassung aufgrund der A-priori-Wahrscheinlichkeiten und Fehlklassifizierungskosten fehlerhaft klassifiziert wurden.
- Bei metrischen abhängigen Variablen ist die Risikoschätzung die Varianz innerhalb der Knoten.

Klassifikationsmatrix. Bei kategorialen (nominalen, ordinalen) abhängigen Variablen zeigt diese Tabelle die Anzahl der Fälle in jeder Kategorie der abhängigen Kategorie, die korrekt bzw. fehlerhaft klassifiziert wurden. Nicht verfügbar für metrische abhängige Variablen.

Abbildung 1-22
Tabellen für Risiko und Klassifizierung

Risiko

Schätzer	Standardfehler
,205	,008

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Klassifikation

	Vorhergesagt		
Beobachtet	schlecht	gut	Prozent korrekt
schlecht	665	355	65,2%
gut	149	1295	89,7%
Gesamtprozentsatz	33,0%	67,0%	79,5%

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Kostenwerte, Werte für A-priori-Wahrscheinlichkeiten, Score-Werte und Profitwerte. Bei kategorialen abhängigen Variablen zeigt diese Tabelle die Kostenwerte, die Werte für die A-priori-Wahrscheinlichkeiten, die Score-Werte und die Profitwerte für die Analyse. Nicht verfügbar für metrische abhängige Variablen.

Unabhängige Variablen

Wichtigkeit für Modell. Bei der CRT-Aufbaumethode wird jede unabhängige Variable (Einflussvariable) gemäß ihrer Bedeutung für das Modell in eine Rangliste eingeordnet. Nicht verfügbar für QUEST- und CHAID-Methoden.

Surrogate nach Aufteilung. Bei den Aufbaumethoden CRT und QUEST werden die Surrogate für jede Aufteilung im Baum aufgeführt, sofern das Modell überhaupt Surrogate enthält. Nicht verfügbar für CHAID-Methoden. Für weitere Informationen siehe Surrogate auf S. 16.

Knotenleistung

Zusammenfassung. Bei metrischen abhängigen Variablen enthält die Tabelle die Knotennummer, die Anzahl der Fälle und den Mittelwert für die abhängige Variable. Bei kategorialen abhängigen Variablen mit definierten Profiten zeigt die Tabelle die Knotennummer, die Anzahl der Fälle, den durchschnittlichen Profit sowie den Anlageertrag (ROI). Nicht verfügbar für kategoriale abhängige Variablen, bei denen keine Profite definiert sind. Für weitere Informationen siehe Profite auf S. 18.

Abbildung 1-23
Gewinnauswertungstabellen für Knoten und Perzentile

Gewinnzusammenfassung für Knoten

Knoten	N	Perzentile	Profit	ROI
7	322	13,1%	77,826	377,4%
5	390	15,8%	70,308	308,8%
6	455	18,5%	67,692	287,9%
9	483	19,6%	49,420	172,0%
8	261	10,6%	23,410	64,7%
1	553	22,4%	22,532	61,9%

Gewinnzusammenfassung für Perzentile

Perzentile	Knoten	N	Profit	ROI
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
50	6 ; 9	1232	70,205	307,9%
60	9	1478	66,745	280,6%
70	9 ; 8	1725	63,134	254,4%
80	8 ; 1	1971	58,149	221,6%
90	1	2218	54,183	197,9%
100	1	2464	51,023	180,4%

Nach Zielkategorie. Bei kategorialen abhängigen Variablen mit definierten Zielkategorien enthält die Tabelle den prozentualen Gewinn, die Antworten in Prozent sowie den Indexprozentsatz (Anhebung) für die einzelnen Knoten- oder Perzentilgruppen. Für jede Zielkategorie wird eine separate Tabelle erstellt. Nicht verfügbar für metrische abhängige Variablen und kategoriale abhängige Variablen, bei denen jeweils keine Zielkategorien definiert sind. Für weitere Informationen siehe Auswählen von Kategorien auf S. 6.

Abbildung 1-24
Zielkategoriegewinne für Knoten und Perzentile

Zielkategorie: Schlecht

Gewinne für Knoten

	Knoten		Gewinn			
Knoten	N	Prozent	N	Prozent	Treffer	Index
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Gewinne für Perzentile

			Gewinn			
Perzentile	Knoten	N	N	Prozent	Treffer	Index
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	604	59,3%	81,8%	197,6%
40	8 ; 9	986	740	72,6%	75,1%	181,3%
50	9	1232	848	83,1%	68,8%	166,2%
60	9 ; 6	1478	908	89,0%	61,4%	148,4%
70	6	1725	951	93,3%	55,1%	133,2%
80	6 ; 5	1971	986	96,7%	50,0%	120,9%
90	5 ; 7	2218	1012	99,3%	45,6%	110,3%
100	7	2464	1020	100,0%	41,4%	100,0%

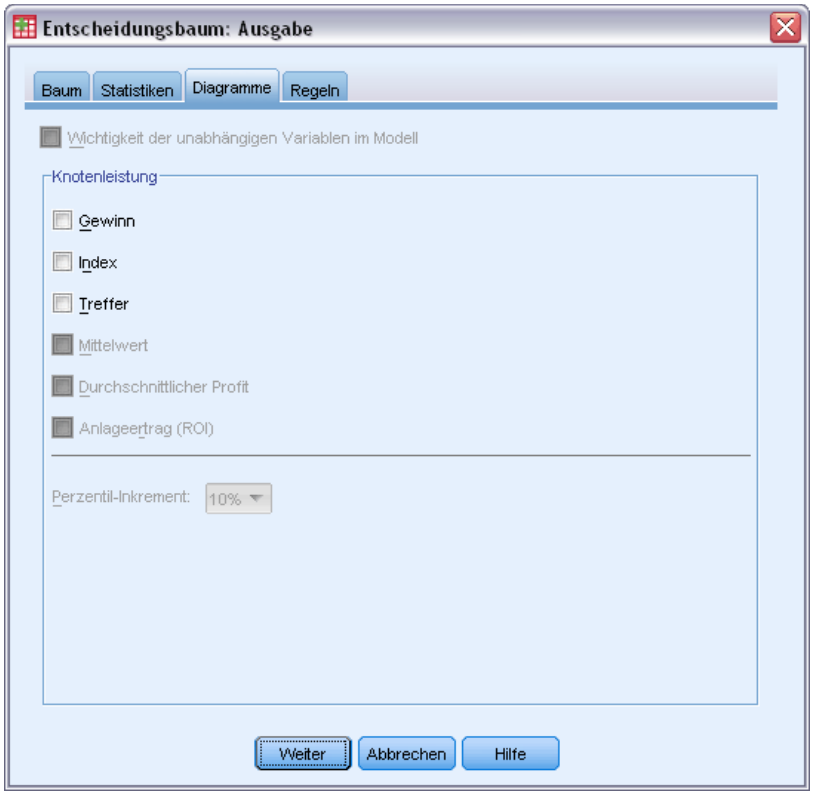
Zeilen. Die Tabellen mit der Knotenleistung können Ergebnisse nach Endknoten und/oder nach Perzentilen aufnehmen. Wenn Sie beide Elemente auswählen, werden je zwei Tabellen für jede Zielkategorie angelegt. Die Perzentiltabellen zeigen kumulative Werte für die einzelnen Perzentile auf der Grundlage der Sortierreihenfolge.

Perzentil-Inkrement. Bei Perzentiltabellen können Sie das Perzentil-Inkrement auswählen: 1, 2, 5, 10, 20 oder 25.

Kumulative Statistik anzeigen. Bei Endknotentabellen werden zusätzliche Spalten mit kumulativen Ergebnissen in die einzelnen Tabellen aufgenommen.

Diagramme

Abbildung 1-25
Dialogfeld "Ausgabe", Registerkarte "Diagramme"



Die verfügbaren Diagramme sind abhängig vom Messniveau der abhängigen Variable, von der Aufbaumethode und anderen Einstellungen.

Wichtigkeit der unabhängigen Variablen im Modell. Balkendiagramm über die Modellbedeutung nach unabhängiger Variable (Einflussvariable). Nur für die CRT-Aufbaumethode verfügbar.

Knotenleistung

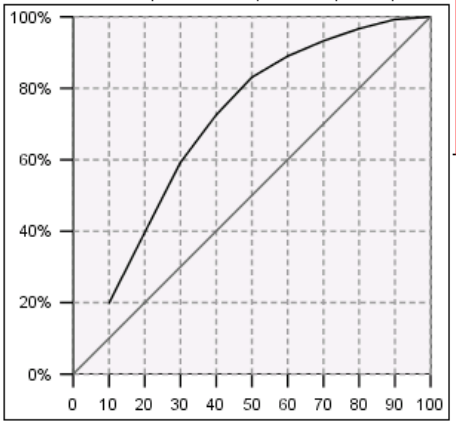
Gewinn. Der Gewinn ist der Prozentsatz aller Fälle in der Zielkategorie in jedem Knoten und wird wie folgt berechnet: (Knotenziel-*n*/Gesamtziel-*n*) x 100. Das Gewinndiagramm besteht aus einem Liniendiagramm kumulativer Perzentilgewinne, die wie folgt berechnet werden: (Kumulatives Perzentilziel-*n*/Gesamtziel-*n*) x 100. Für jede Zielkategorie wird ein separates Liniendiagramm erstellt. Nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind. Für weitere Informationen siehe Auswählen von Kategorien auf S. 6.

Das Gewinndiagramm enthält dieselben Werte wie die Spalte *Gewinn (Prozent)* in der Tabelle "Gewinne für Perzentile"; hier werden ebenfalls kumulative Werte angezeigt.

Abbildung 1-26
Tabelle "Gewinne für Perzentile" und Gewinndiagramm

Gewinne für Perzentile

Perzentile	Knoten	N	Gewinn N	Gewinn Prozent	Treffer	Index
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	604	59,3%	81,8%	197,6%
40	8 ; 9	986	740	72,6%	75,1%	181,3%
				83,1%	68,8%	166,2%
				89,0%	61,4%	148,4%
				93,3%	55,1%	133,2%
				96,7%	50,0%	120,9%
				99,3%	45,6%	110,3%
				100,0%	41,4%	100,0%



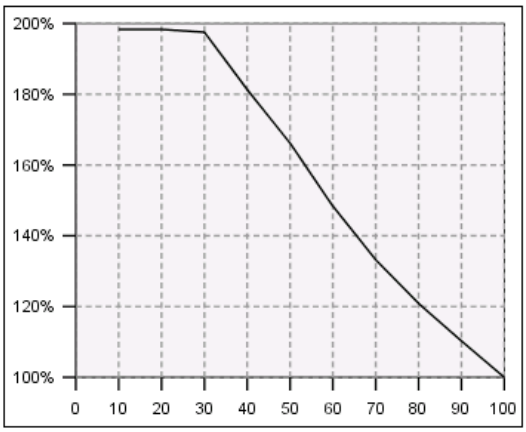
Index. Der Index ist das Verhältnis des Zielkategorieanteils im Knoten zum Zielkategorieanteil der gesamten Stichprobe. Das Indexdiagramm ist ein Liniendiagramm kumulativer Perzentil-Indexwerte. Nur für kategoriale abhängige Variablen verfügbar. Der kumulative Perzentil-Index wird wie folgt berechnet: (Kumulative Perzentil-Antwort in Prozent/Gesamtantwort in Prozent) x 100. Für jede Zielkategorie wird ein separates Diagramm angelegt. Die Zielkategorien müssen definiert werden.

Das Indexdiagramm enthält dieselben Werte wie die Spalte *Index* in der Tabelle "Gewinne für Perzentile".

Abbildung 1-27
Tabelle "Gewinne für Perzentile" und Indexdiagramm

Gewinne für Perzentile

Perzentile	Knoten	N	Gewinn N	Gewinn Prozent	Treffer	Index
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	604	59,3%	81,8%	197,6%
40	8 ; 9	986	740	72.6%	75,1%	181,3%
						166,2%
						148,4%
						133,2%
						120,9%
						110,3%
						100,0%



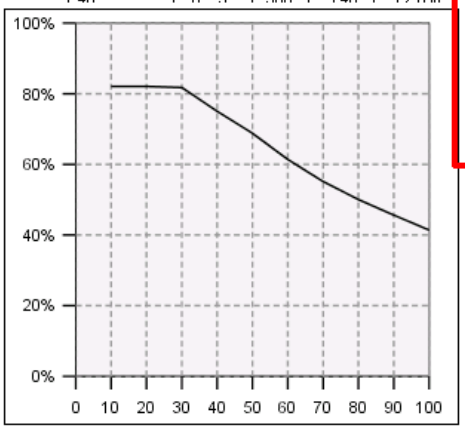
Zielkategorie. Der Prozentsatz der Fälle im Knoten, die der Zielkategorie angehören. Das Antwortdiagramm besteht aus einem Liniendiagramm kumulativer Perzentil-Antworten, die wie folgt berechnet werden: (Kumulatives Perzentilziel-*n*/Kumulatives Perzentil-Gesamt-*n*) x 100. Dies ist nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind.

Das Antwortdiagramm enthält dieselben Werte wie die Spalte *Antwort* in der Tabelle "Gewinne für Perzentile".

Abbildung 1-28
Tabelle "Gewinne für Perzentile" und Antwortdiagramm

Gewinne für Perzentile

Perzentile	Knoten	N	Gewinn N	Gewinn Prozent	Treffer	Index
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	604	59,3%	81,8%	197,6%
40	8 · 9	986	740	72,6%	75,1%	181,3%
					68,8%	166,2%
					61,4%	148,4%
					55,1%	133,2%
					50,0%	120,9%
					45,6%	110,3%
					41,4%	100,0%



Mittelwert. Liniendiagramm der kumulativen Perzentil-Mittelwerte für die abhängige Variable. Nur für metrische abhängige Variablen verfügbar.

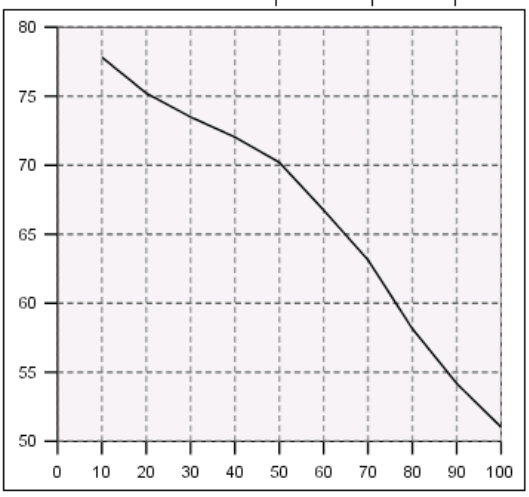
Durchschnittlicher Profit. Liniendiagramm des kumulativen durchschnittlichen Profits. Nur für kategoriale abhängige Variablen verfügbar, bei denen Profite definiert sind. Für weitere Informationen siehe Profite auf S. 18.

Das Diagramm für den durchschnittlichen Profit enthält dieselben Werte wie die Spalte *Profit* in der Tabelle "Gewinnzusammenfassung für Perzentile".

Abbildung 1-29
Tabelle "Gewinnzusammenfassung für Perzentile" und Durchschnittsprofit-Diagramm

Gewinnzusammenfassung für Perzentile

Perzentile	Knoten	N	Profit	ROI
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
			70,205	307,9%
			66,745	280,6%
			63,134	254,4%
			58,149	221,6%
			54,183	197,9%
			51,023	180,4%



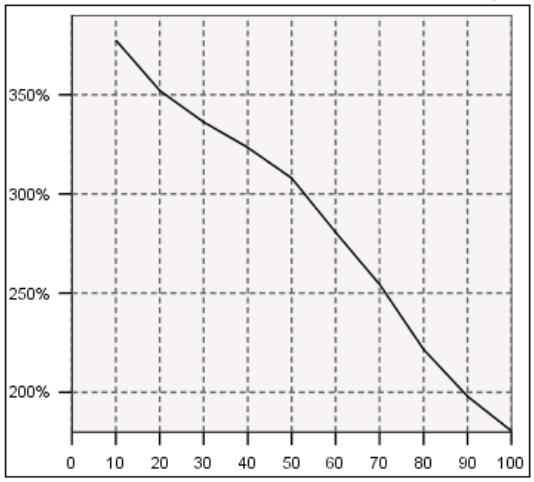
Anlageertrag (ROI). Liniendiagramm des kumulativen ROI (Anlageertrag). Der ROI wird als Verhältnis der Profite zu den Aufwendungen berechnet. Nur für kategoriale abhängige Variablen verfügbar, bei denen Profite definiert sind.

Das ROI-Diagramm enthält dieselben Werte wie die Spalte *ROI* in der Tabelle "Gewinnzusammenfassung für Perzentile".

Abbildung 1-30
Tabelle "Gewinnzusammenfassung für Perzentile" und ROI-Diagramm

Gewinnzusammenfassung für Perzentile

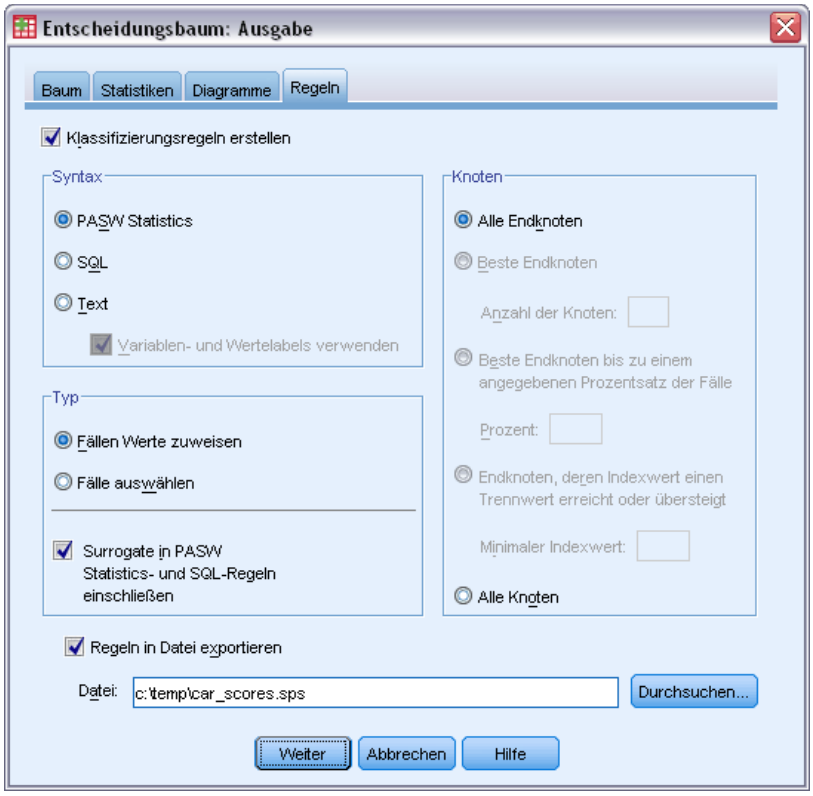
Perzentile	Knoten	N	Profit	ROI
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
				307,9%
				280,6%
				254,4%
				221,6%
				197,9%
				180,4%



Perzentil-Inkrement. Bei allen Perzentildiagrammen steuert diese Einstellung die im Diagramm abgebildeten Perzentil-Inkremente: 1, 2, 5, 10, 20 oder 25.

Auswahl- und Bewertungsregeln

Abbildung 1-31
Dialogfeld "Ausgabe", Registerkarte "Regeln"



Auf der Registerkarte "Regeln" legen Sie die Regeln für die Auswahl oder die Klassifizierung/Vorhersage mit der Befehlssyntax, als SQL-Anweisungen oder in natürlicher Sprache fest. Sie können diese Regeln im Viewer anzeigen lassen und/oder in einer externen Datei speichern.

Syntax. Steuert die Form der Auswahlregeln sowohl für die Ausgabe im Viewer als auch beim Speichern in einer externen Datei.

- **PASW Statistics.** Befehlssyntax-Sprache. Die Regeln werden als Befehle ausgedrückt, die eine Filterbedingung zum Auswählen von Untergruppen mit Fällen definieren, oder auch als `COMPUTE`-Anweisungen, mit denen Fälle bewertet werden können.
- **SQL.** Um Datensätze auszuwählen oder aus einer Datenbank zu extrahieren oder um Werte für diese Datensätze zuzuweisen, werden Standard-SQL-Regeln erzeugt. Die erzeugten SQL-Regeln enthalten keine Tabellennamen oder andere Informationen zur Datenquelle.
- **Text.** Pseudo-Code in natürlicher Sprache. Regeln werden als Reihe logischer Wenn-Dann-Anweisungen ausgedrückt, die die Klassifizierungen oder Vorhersagen des Modells für jeden Knoten beschreiben. Regeln in dieser Form können definierte Variablen- und Wertelabels oder auch Variablennamen und Datenwerte nutzen.

Typ. Bei PASW Statistics- und SQL-Regeln wird hiermit der Typ der erzeugten Regeln gesteuert: Auswahl- oder Bewertungsregeln.

- **Fällen Werte zuweisen.** Mit den Regeln können die Vorhersagen aus dem Modell Fällen zugewiesen werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Für jeden Knoten, der den Kriterien für die Knotenzugehörigkeit entspricht, wird eine separate Regel erzeugt.
- **Fälle auswählen.** Mit den Regeln können Fälle ausgewählt werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Bei PASW Statistics- und SQL-Regeln wird eine einzige Regel erzeugt, mit der alle Fälle ausgewählt werden, die den Auswahlkriterien entsprechen.

Ersatzwerte in PASW Statistics- und SQL-Regeln einschließen. Bei CRT und QUEST können Sie ersatzweise Einflussvariablen aus dem Modell in die Regeln aufnehmen. Regeln mit Surrogaten können recht komplex werden. Wenn Sie nur konzeptuelle Daten zu Ihrem Baum ableiten möchten, sollten Sie die Surrogate ausschließen. Wenn die Daten in den unabhängigen Variablen (Einflussvariablen) in bestimmten Fällen unvollständig sind und Regeln angelegt werden sollen, die den Baum getreu nachbilden, schließen Sie die Surrogate ein. Für weitere Informationen siehe Surrogate auf S. 16.

Knoten. Steuert den Umfang der erzeugten Regeln. Für jeden Knoten im Umfang wird eine separate Regel erzeugt.

- **Alle Endknoten.** Erzeugt Regeln für jeden Endknoten.
- **Beste Endknoten.** Erzeugt Regeln für die besten *n* Endknoten auf der Grundlage der Indexwerte. Ist die Anzahl höher als die Anzahl der Endknoten im Baum, werden Regeln für alle Endknoten erzeugt. (Siehe nachstehende Anmerkung.)
- **Beste Endknoten bis zu einem angegebenen Prozentsatz der Fälle.** Erzeugt Regeln für Endknoten für die oberen *n* Prozent der Fälle auf der Grundlage der Indexwerte. (Siehe nachstehende Anmerkung.)
- **Endknoten, deren Indexwert einen Trennwert erreicht oder übersteigt.** Erzeugt Regeln für alle Endknoten, deren Indexwert größer oder gleich dem angegebenen Wert ist. Ein Indexwert größer als 100 bedeutet, dass der Prozentsatz der Fälle in der Zielkategorie in diesem Knoten größer ist als der Prozentsatz im Stammknoten. (Siehe nachstehende Anmerkung.)
- **Alle Knoten.** Erzeugt Regeln für alle Knoten.

Anmerkung 1: Die Knotenauswahl auf der Grundlage der Indexwerte ist nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind. Wenn Sie mehrere Zielkategorien angegeben haben, wird je ein Regelsatz für die einzelnen Zielkategorien erzeugt.

Anmerkung 2: Bei PASW Statistics- und SQL-Regeln zum Auswählen von Fällen (nicht bei Regeln zum Zuweisen von Werten) wird mit den Optionen Alle Knoten und Alle Endknoten eine Regel erzeugt, mit der alle Fälle in der Analyse ausgewählt werden.

Regeln in Datei exportieren. Speichert die Regeln in einer externen Textdatei.

Alternativ können Sie die Auswahl- und Bewertungsregeln interaktiv anhand ausgewählter Knoten im fertigen Baummodell erzeugen und speichern. Für weitere Informationen siehe Regeln für die Auswahl oder Bewertung von Fällen in Kapitel 2 auf S. 49.

Hinweis: Wenn Sie Regeln als Befehlssyntax auf eine andere Datendatei anwenden, müssen die Namen der Variablen in dieser Datendatei mit den Namen der unabhängigen Variablen im fertigen Modell identisch sein. Des Weiteren müssen die Variablen mit derselben Maßeinheit gemessen werden und dieselben benutzerdefiniert fehlenden Werte aufweisen (falls vorhanden).

Kapitel
2

Baumeditor

Der Baumeditor bietet die folgenden Möglichkeiten:

- Ausgewählte Baumverzweigungen ein- und ausblenden.
- Anzeige des Knoteninhalts, der Statistiken an den Knotenaufteilungen und anderer Informationen steuern.
- Farben für Knoten, Hintergrund, Rahmen, Diagramme und Schriften ändern.
- Schriftart und -größe ändern.
- Baumausrichtung ändern.
- Untergruppen von Fällen für weitere Analyse auf der Grundlage ausgewählter Knoten auswählen.
- Regeln zum Auswählen und Bewerten von Fällen auf der Grundlage ausgewählter Knoten erstellen und speichern.

So bearbeiten Sie ein Baummodell:

▶ Doppelklicken Sie im Viewer-Fenster auf das Baummodell.

oder

▶ Wählen Sie im Menü “Bearbeiten” bzw. im Kontextmenü folgende Optionen:

Inhalt bearbeiten
In separatem Fenster

Ein- und Ausblenden von Knoten

So können Sie alle untergeordneten Knoten in einer Verzweigung unterhalb eines übergeordneten Knotens ausblenden (reduzieren):

▶ Klicken Sie auf das Minuszeichen (–) in dem kleinen Kästchen unterhalb der rechten unteren Ecke des übergeordneten Knotens.

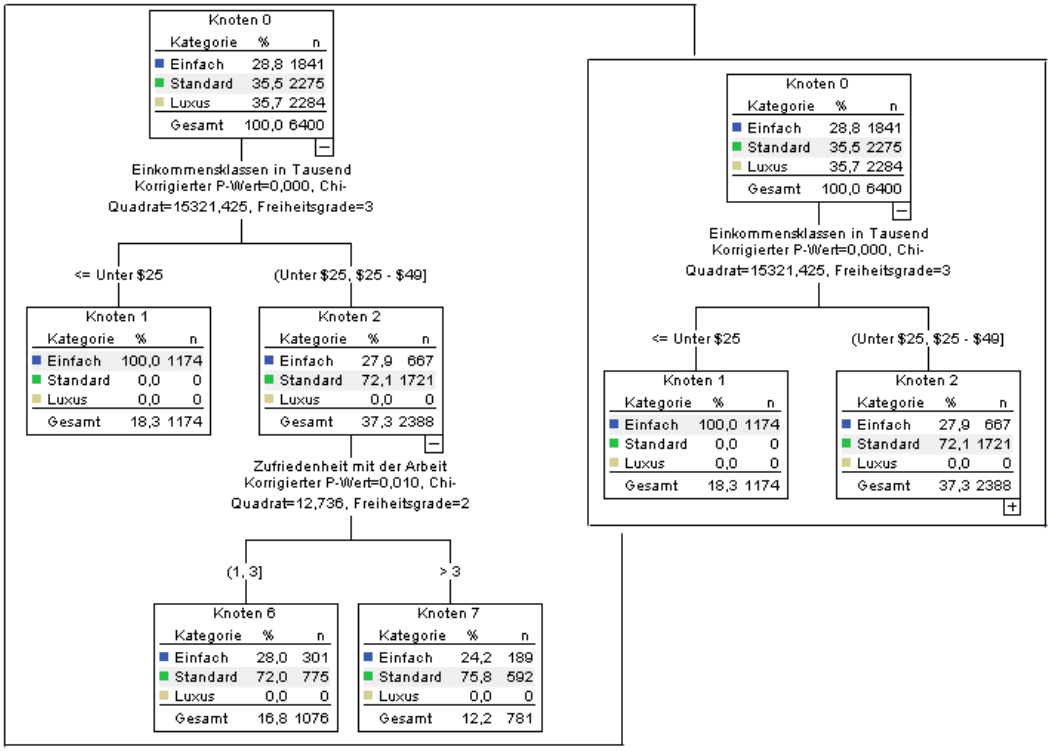
Alle Knoten unterhalb des übergeordneten Knotens in dieser Verzweigung werden ausgeblendet.

So können Sie die untergeordneten Knoten in einer Verzweigung unterhalb eines übergeordneten Knotens einblenden (erweitern):

▶ Klicken Sie auf das Pluszeichen () in dem kleinen Kästchen unterhalb der unteren rechten Ecke des übergeordneten Knotens.

Hinweis: Das Ausblenden der untergeordneten Knoten in einer Verzweigung ist nicht dasselbe wie das Beschneiden eines Baums. Soll der Baum beschnitten werden, aktivieren Sie das Beschneiden, bevor Sie den Baum erstellen. Beschnittene Verzweigungen sind nicht im endgültigen Baum enthalten. Für weitere Informationen siehe Beschneiden von Bäumen in Kapitel 1 auf S. 15.

Abbildung 2-1
Erweiterter und reduzierter Baum



Auswählen mehrerer Knoten

Auf der Grundlage des oder der ausgewählten Knoten können Sie Fälle auswählen, Bewertungs- und Auswahlregeln erstellen und andere Aktionen ausführen. So wählen Sie mehrere Knoten aus:

- Klicken Sie auf einen Knoten.
- Halten Sie die STRG-Taste gedrückt und klicken Sie auf die weiteren Knoten.

Sie können mehrere Knoten auf derselben Ebene und/oder übergeordnete Knoten in einer Verzweigung auswählen und untergeordnete Knoten in einer anderen Verzweigung. Es ist allerdings nicht möglich, gleichzeitig einen übergeordneten Knoten und einen untergeordneten Knoten bzw. einen Nachfolger in derselben Knotenverzweigung auszuwählen.

Arbeiten mit umfangreichen Bäumen

Baummodelle enthalten manchmal so viele Knoten und Verzweigungen, dass der gesamte Baum nur schwer oder auch gar nicht vollständig und in der vollen Größe angezeigt werden kann. Beim Arbeiten mit umfangreichen Bäumen steht eine Reihe nützlicher Funktionen bereit:

- **Baumstruktur.** Mithilfe der Baumstruktur, eine stark verkleinerte, vereinfachte Version des Baums, können Sie im Baum navigieren und Knoten auswählen. Für weitere Informationen siehe Baumstruktur auf S. 43.
- **Skalierung.** Zum Vergrößern und Verkleinern ändern Sie den Skalierungsprozentsatz für die Baumanzeige. Für weitere Informationen siehe Skalieren der Baumanzeige auf S. 44.
- **Knoten- und Verzweigungsanzeige.** Um einen Baum kompakter zu gestalten, können Sie nur Tabellen oder nur Diagramme in den Knoten anzeigen lassen und/oder die Anzeige von Knotenbeschriftungen oder Informationen zu unabhängigen Variablen unterdrücken. Für weitere Informationen siehe Steuern der im Baum angezeigten Daten auf S. 46.

Baumstruktur

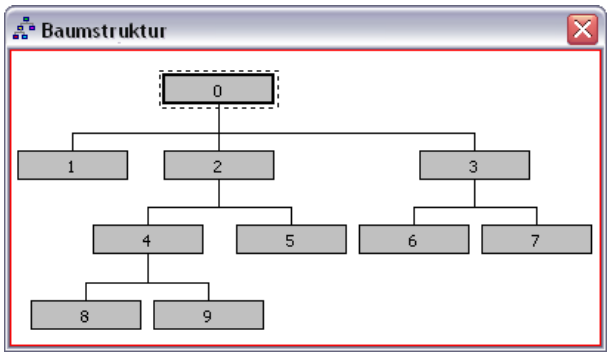
Die Baumstruktur ist eine kompakte, vereinfachte Ansicht des Baums, mit der Sie im Baum navigieren und Knoten auswählen können.

So verwenden Sie das Baumstruktur-Fenster:

▶ Wählen Sie die folgenden Menübefehle des Baum-Editors aus:

Ansicht
Baumstruktur

Abbildung 2-2
Baumstruktur-Fenster



- Der derzeit ausgewählte Knoten ist sowohl im Baummodell-Editor als auch im Baumstruktur-Fenster hervorgehoben.
- Der Teil des Baums, der derzeit im Ansichtsbereich des Baummodell-Editors angezeigt wird, ist in der Baumstruktur mit einem roten Rechteck umrandet. Soll ein anderer Teil des Baums im Ansichtsbereich dargestellt werden, klicken Sie mit der rechten Maustaste auf das Rechteck und ziehen Sie es an die gewünschte Position.

- Wenn Sie einen Knoten in der Baumstruktur auswählen, der sich derzeit im Ansichtsbereich des Baumeditors befindet, wird der sichtbare Ausschnitt so verschoben, dass der ausgewählte Knoten sichtbar wird.
- Die Mehrfachknotenauswahl funktioniert in der Baumstruktur auf dieselbe Weise wie im Baumeditor: Halten Sie die STRG-Taste gedrückt und wählen Sie die gewünschten Knoten aus. Es ist nicht möglich, gleichzeitig einen übergeordneten Knoten und einen untergeordneten Knoten bzw. einen Nachfolger in derselben Knotenverzweigung auszuwählen.

Skalieren der Baumanzeige

Standardmäßig werden Bäume so skaliert, dass sie vollständig im Viewer-Fenster dargestellt werden können. Bei bestimmten Bäumen sind die Angaben daher unter Umständen nur schwer lesbar. Wählen Sie eine vordefinierte Einstellung für die Skalierung aus oder geben Sie einen benutzerdefinierten Wert zwischen 5 % und 200 % ein.

So ändern Sie die Skalierung des Baums:

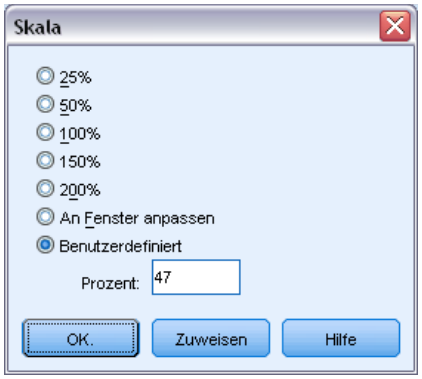
▶ Wählen Sie einen Skalierungsprozentsatz in der Dropdown-Liste in der Symbolleiste aus oder geben Sie einen benutzerdefinierten Wert ein.

oder

▶ Wählen Sie die folgenden Menübefehle des Baum-Editors aus:

Ansicht
Skala...

Abbildung 2-3
Dialogfeld "Skala"



Des Weiteren können Sie einen Skalierungswert angeben, noch bevor Sie das Baummodell erstellen. Für weitere Informationen siehe Ausgabe in Kapitel 1 auf S. 25.

Knotenübersichtsfenster

Das Knotenübersichtsfenster ermöglicht einen genaueren Blick auf die ausgewählten Knoten. Im Übersichtsfenster können Sie außerdem Auswahl- und Bewertungsregeln auf der Grundlage der ausgewählten Knoten anzeigen lassen, anwenden und speichern.

- Mit dem Menü “Ansicht” im Knotenübersichtsfenster wechseln Sie zwischen einer Übersichtstabelle, einem Diagramm und den Regeln.
- Im Menü “Regeln” im Knotenübersichtsfenster wählen Sie den Typ für die anzuzeigenden Regeln aus. Für weitere Informationen siehe Regeln für die Auswahl oder Bewertung von Fällen auf S. 49.
- Alle Ansichten im Knotenübersichtsfenster zeigen eine kombinierte Übersicht für alle ausgewählten Knoten.

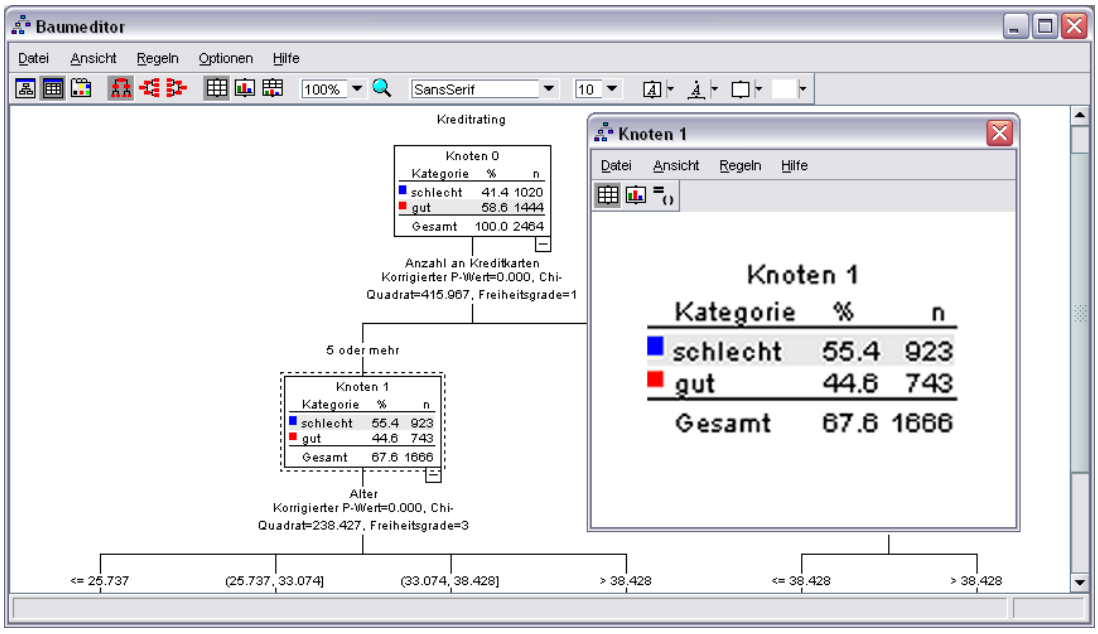
So verwenden Sie das Knotenübersichtsfenster:

▶ Wählen Sie die gewünschten Knoten im Baumeditor aus. Sollen mehrere Knoten ausgewählt werden, halten Sie beim Klicken die STRG-Taste gedrückt.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Ansicht
  Zusammenfassung

Abbildung 2-4
Übersichtsfenster

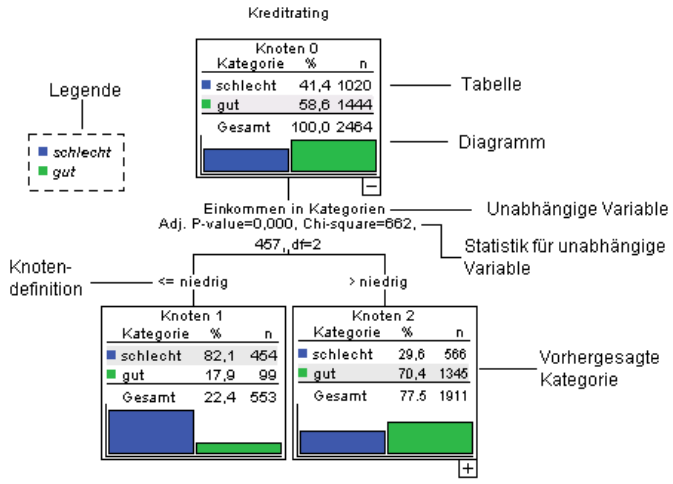


Steuern der im Baum angezeigten Daten

Mit dem Menü “Optionen” im Baumeditor steuern Sie die Anzeige des Knoteninhalts, der Namen und Statistiken der unabhängigen Variablen (Einflussvariablen), der Knotendefinitionen und andere Einstellungen. Ein Großteil der Einstellungen kann auch über die Symbolleiste gesteuert werden.

Einstellung	**Auswahl im Menü “Optionen”**
Vorhergesagte Kategorie hervorheben (kategoriale abhängige Variable)	Vorhergesagten Wert hervorheben
Tabellen und/oder Diagramme in Knoten	Knoteninhalt
Signifikanztestwerte und *p*-Werte	Statistik für unabhängige Variablen
Namen von unabhängigen Variablen (Einflussvariablen)	Unabhängige Variablen
Unabhängige(r) Wert(e) (Einflusswert[e]) für Knoten	Knotendefinitionen
Ausrichtung (von oben nach unten, von links nach rechts, von rechts nach links)	Ausrichtung
Diagrammlegende	Legende

Abbildung 2-5
Baumelemente



Ändern der Farben und Schriftarten im Baum

Die folgenden Farben im Baum können geändert werden:

- Rahmen-, Hintergrund- und Textfarbe für Knoten
- Farbe und Textfarbe für Verzweigungen
- Farbe für den Baumhintergrund
- Hervorhebungsfarbe für vorhergesagte Kategorien (kategoriale abhängige Variablen)
- Farben in Knotendiagrammen

Des Weiteren können Sie die Schriftart, den Schriftschnitt und die Schriftgröße für den gesamten Text im Baum ändern.

Hinweis: Es ist nicht möglich, die Farbe oder die Schriftattribute für einzelne Knoten oder Verzweigungen zu ändern. Farbänderungen gelten für sämtliche Elemente desselben Typs, Änderungen an der Schriftart (mit Ausnahme der Farben) gelten für alle Diagrammelemente.

So ändern Sie die Farben und die Schriftattribute:

- Ändern Sie die Schriftattribute für den gesamten Baum bzw. die Farben für verschiedene Elemente über die Symbolleiste. (Wenn Sie mit der Maus auf eine Steuerung in der Symbolleiste zeigen, wird eine QuickInfo mit einer Beschreibung für diese Steuerung eingeblendet.)

  oder

- Öffnen Sie das Fenster "Eigenschaften". Doppelklicken Sie hierzu auf eine beliebige Stelle im Baumeditor oder wählen Sie die folgenden Befehle aus den Menüs aus:

  Ansicht
  Eigenschaften

- Rahmen, Verzweigung, Knotenhintergrund, vorhergesagte Kategorie, Baumhintergrund: Klicken Sie auf die Registerkarte Farbe.

- Schriftfarbe und Schriftattribute: Klicken Sie auf die Registerkarte Text.

- Farben in Knotendiagrammen: Klicken Sie auf die Registerkarte Knotendiagramme.

Abbildung 2-6
Fenster "Eigenschaften", Registerkarte "Farbe"

Abbildung 2-7
Fenster "Eigenschaften", Registerkarte "Text"

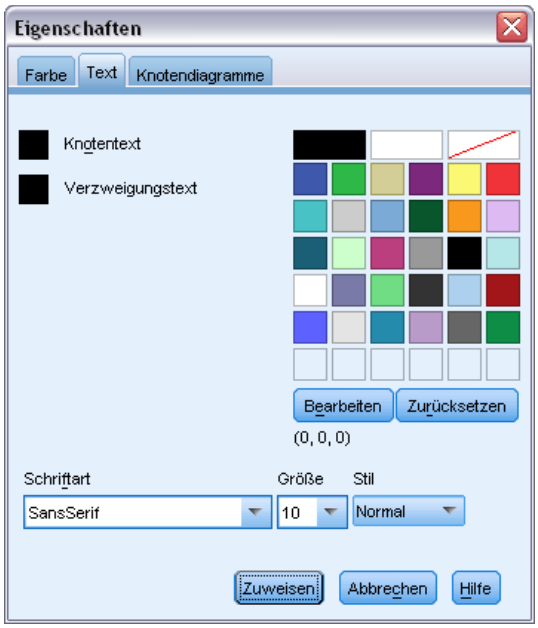
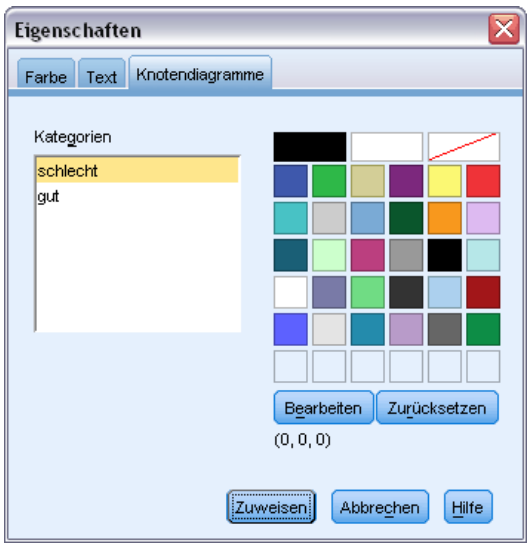


Abbildung 2-8
Fenster "Eigenschaften", Registerkarte "Knotendiagramme"



Regeln für die Auswahl oder Bewertung von Fällen

Der Baumeditor bietet die folgenden Möglichkeiten:

- Teilgruppen von Fällen auf der Grundlage des oder der ausgewählten Knoten auswählen. Für weitere Informationen siehe Filtern von Fällen auf S. 49.
- Regeln für die Auswahl oder Bewertung von Fällen im PASW Statistics- oder SQL-Format erzeugen. Für weitere Informationen siehe Speichern von Auswahl- und Bewertungsregeln auf S. 49.

Wenn Sie das Baummodell mit der Prozedur “Entscheidungsbaum” erstellen, können Sie außerdem die Regeln automatisch nach bestimmten Kriterien speichern lassen. Für weitere Informationen siehe Auswahl- und Bewertungsregeln in Kapitel 1 auf S. 38.

Filtern von Fällen

Wenn Sie weitere Informationen zu den Fällen in einem bestimmten Knoten oder einer Knotengruppe benötigen, können Sie eine Untergruppe mit Fällen für die weitere Analyse auf der Grundlage der ausgewählten Knoten auswählen.

▶ Wählen Sie die gewünschten Knoten im Baumeditor aus. Sollen mehrere Knoten ausgewählt werden, halten Sie beim Klicken die STRG-Taste gedrückt.

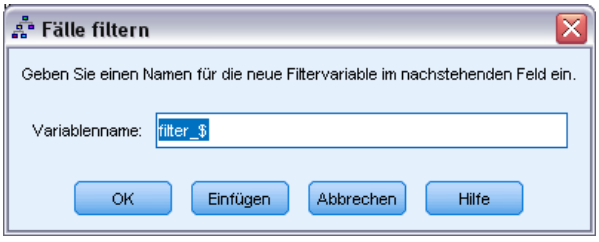
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Regeln
  Fälle filtern...

▶ Geben Sie einen Namen für die Filtervariable an. Die Fälle aus den ausgewählten Knoten erhalten den Wert 1 für diese Variable. Alle anderen Fälle erhalten den Wert 0 und werden aus der weiteren Analyse ausgeschlossen, bis der Filterstatus geändert wird.

▶ Klicken Sie auf OK.

Abbildung 2-9
Dialogfeld “Fälle filtern”



Speichern von Auswahl- und Bewertungsregeln

Sie können die Auswahl- und Bewertungsregeln in einer externen Datei speichern und dann auf eine andere Datenquelle anwenden. Die Regeln beruhen auf den ausgewählten Knoten im Baumeditor.

Syntax. Steuert die Form der Auswahlregeln sowohl für die Ausgabe im Viewer als auch beim Speichern in einer externen Datei.

- **PASW Statistics.** Befehlssyntax-Sprache. Die Regeln werden als Befehle ausgedrückt, die eine Filterbedingung zum Auswählen von Untergruppen mit Fällen definieren, oder auch als COMPUTE-Anweisungen, mit denen Fälle bewertet werden können.
- **SQL.** Um Datensätze auszuwählen oder aus einer Datenbank zu extrahieren oder um Werte für diese Datensätze zuzuweisen, werden Standard-SQL-Regeln erzeugt. Die erzeugten SQL-Regeln enthalten keine Tabellennamen oder andere Informationen zur Datenquelle.

Typ. Sie können Auswahl- oder Bewertungsregeln erstellen.

- **Fälle auswählen.** Mit den Regeln können Fälle ausgewählt werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Bei PASW Statistics- und SQL-Regeln wird eine einzige Regel erzeugt, mit der alle Fälle ausgewählt werden, die den Auswahlkriterien entsprechen.
- **Fällen Werte zuweisen.** Mit den Regeln können die Vorhersagen aus dem Modell Fällen zugewiesen werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Für jeden Knoten, der den Kriterien für die Knotenzugehörigkeit entspricht, wird eine separate Regel erzeugt.

Ersatzwerte berücksichtigen. Bei CRT und QUEST können Sie ersatzweise Einflussvariablen aus dem Modell in die Regeln aufnehmen. Regeln mit Surrogaten können recht komplex werden. Wenn Sie nur konzeptuelle Daten zu Ihrem Baum ableiten möchten, sollten Sie die Surrogate ausschließen. Wenn die Daten in den unabhängigen Variablen (Einflussvariablen) in bestimmten Fällen unvollständig sind und Regeln angelegt werden sollen, die den Baum getreu nachbilden, schließen Sie die Surrogate ein. Für weitere Informationen siehe Surrogate in Kapitel 1 auf S. 16.

So speichern Sie Auswahl- oder Bewertungsregeln für Fälle:

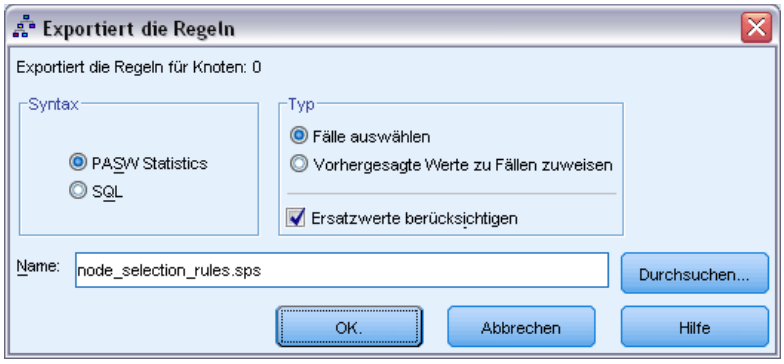
▶ Wählen Sie die gewünschten Knoten im Baumeditor aus. Sollen mehrere Knoten ausgewählt werden, halten Sie beim Klicken die STRG-Taste gedrückt.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Regeln
  Exportieren...

▶ Wählen Sie den gewünschten Regeltyp aus und geben Sie einen Dateinamen ein.

Abbildung 2-10
Dialogfeld "Regeln exportieren"



Hinweis: Wenn Sie Regeln als Befehlssyntax auf eine andere Datendatei anwenden, müssen die Namen der Variablen in dieser Datendatei mit den Namen der unabhängigen Variablen im fertigen Modell identisch sein. Des Weiteren müssen die Variablen mit derselben Maßeinheit gemessen werden und dieselben benutzerdefiniert fehlenden Werte aufweisen (falls vorhanden).

Teil II: Beispiele

Kapitel
3

Datenannahmen und -anforderungen

Die Prozedur “Entscheidungsbaum” geht von folgenden Annahmen aus:

- Allen Analysevariablen wurde das richtige Messniveau zugewiesen.
- Bei kategorialen (**nominalen**, **ordinalen**) abhängigen Variablen wurden für alle Kategorien Wertelabels definiert, die in die Analyse aufgenommen werden sollten.

Wir verwenden die Datei *tree_textdata.sav*, um die Wichtigkeit dieser beiden Anforderungen zu verdeutlichen. Diese Datendatei spiegelt den Standardzustand von eingelesenen oder eingegebenen Daten vor der Definition von Attributen, wie Messniveau oder Wertelabels, wider. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Decision Trees (Entscheidungsbäume) 18*.

Auswirkungen des Messniveaus auf Baummodelle

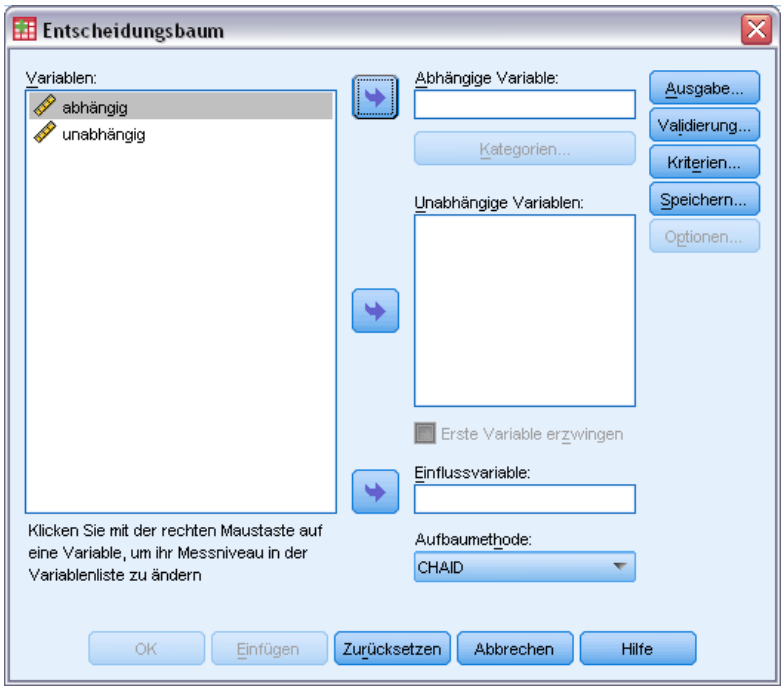
Beide Variablen in dieser Datendatei sind numerisch. Standardmäßig geht man davon aus, dass numerische Variablen ein **metrisches** Messniveau aufweisen. Wie wir jedoch weiter unten sehen werden, handelt es sich bei beiden Variablen in Wahrheit um kategoriale Variablen, bei denen numerische Codes für Kategoriewerte stehen.

▶ Zum Erstellen einer Entscheidungsbaum-Analyse wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Klassifizieren
    Baum...

Die Symbole neben den beiden Variablen in der Quellvariablenliste zeigen an, dass sie als metrische Variablen behandelt werden.

Abbildung 3-1
Hauptdialogfeld von "Entscheidungsbaum" mit zwei metrischen Variablen



- Wählen Sie *abhängig* als abhängige Variable aus.
- Wählen Sie *unabhängig* als unabhängige Variable aus.
- Klicken Sie auf OK, um die Prozedur auszuführen.
- Öffnen Sie noch einmal das Dialogfeld "Entscheidungsbaum" und klicken Sie auf Zurücksetzen.
- Klicken Sie in der Quell-Liste auf *abhängig* und wählen Sie im Kontextmenü die Option Nominal aus.
- Führen Sie denselben Vorgang für die Variable *unabhängig* in der Quell-Liste aus.

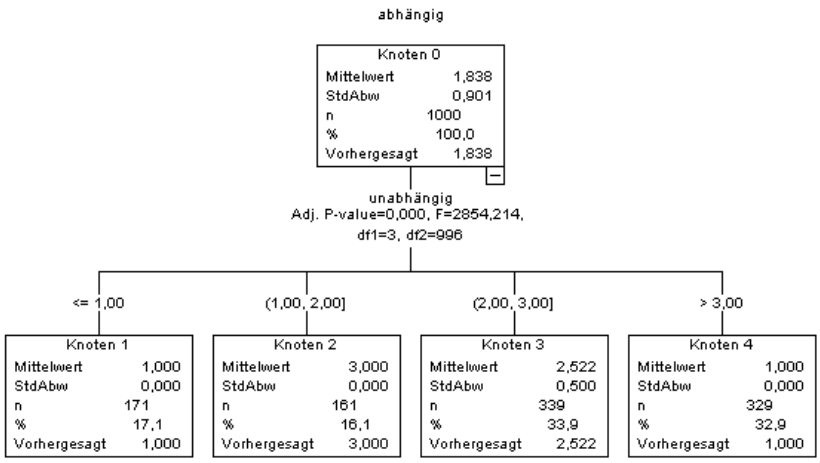
Die Symbole neben den einzelnen Variablen geben nun an, dass sie als nominale Variablen behandelt werden.

Abbildung 3-2
Symbole für "nominal" in der Quell-Liste

▶ Wählen Sie *abhängig* als abhängige Variable und *unabhängig* als unabhängige Variable aus und klicken Sie auf OK, um die Prozedur erneut auszuführen.

Vergleichen wir nun die beiden Bäume. Betrachten wir zunächst den Baum, in dem beide numerischen Variablen als metrische Variablen behandelt werden.

Abbildung 3-3
Baum, bei dem beide Variablen als metrische Variablen behandelt werden

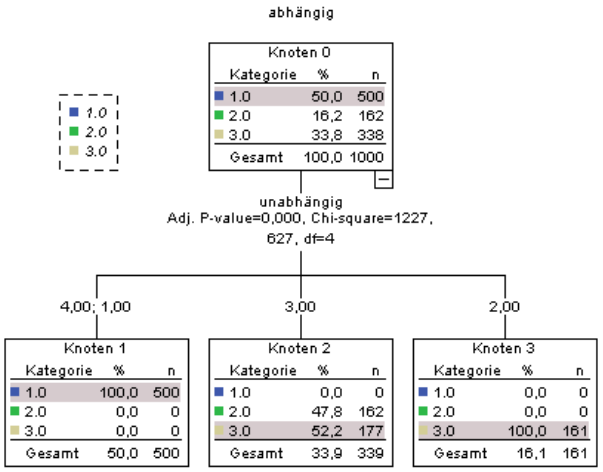


- Jeder Knoten des Baums zeigt den “vorhergesagten” Wert, den Mittelwert für die abhängige Variable an diesem Knoten. Für eine eigentlich kategoriale Variable ist der Mittelwert möglicherweise keine sinnvolle Statistik.
- Der Baum weist vier untergeordnete Knoten auf, einen für jeden Wert der unabhängigen Variablen.

In Baummodellen werden häufig ähnliche Knoten zusammengeführt, bei metrischen Variablen können jedoch nur aufeinanderfolgende Werte zusammengeführt werden. In diesem Beispiel wurden keine aufeinander folgenden Werte als ähnlich genug für eine Knotenzusammenführung betrachtet.

Der Baum, bei dem beide Variablen als nominal behandelt werden, weist in mehrerlei Hinsicht Unterschiede auf.

Abbildung 3-4
Baum, bei dem beide Variablen als nominale Variablen behandelt werden



- Statt eines vorhergesagten Werts enthält jeder Knoten eine Häufigkeitstabelle, die die Anzahl und Prozentsatz der Fälle für jede Kategorie der abhängigen Variablen anzeigt.
- Die “vorhergesagte” Kategorie – die Kategorie mit der höchsten Anzahl in jedem Knoten – ist markiert. Die vorhergesagte Kategorie für Knoten 2 beispielsweise ist Kategorie 3.
- Anstelle von vier untergeordneten Knoten gibt es nur drei, bei denen zwei Werte der unabhängigen Variablen in einen einzelnen Knoten zusammengeführt wurden.

Bei den beiden unabhängigen Werten, die im selben Knoten zusammengeführt wurden, handelt es sich um 1 und 4. Da nominale Werte definitionsgemäß keine natürliche Reihenfolge aufweisen, ist die Zusammenführung nicht aufeinander folgender Werte zulässig.

Dauerhafte Zuweisung des Messniveaus

Wenn Sie das Messniveau für eine Variable im Dialogfeld “Entscheidungsbaum” ändern, gilt diese Änderung nur vorübergehend; sie wird nicht zusammen mit der Datendatei gespeichert. Außerdem ist nicht immer bekannt, was das richtige Messniveau für alle Variablen sein sollte.

Durch “Variableneigenschaften definieren” können Sie das richtige Messniveau für die einzelnen Variablen bestimmen und das zugewiesene Messniveau dauerhaft ändern. So verwenden Sie die Option “Variableneigenschaften definieren”:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
Variableneigenschaften definieren...

Auswirkungen der Wertelabels auf Baummodelle

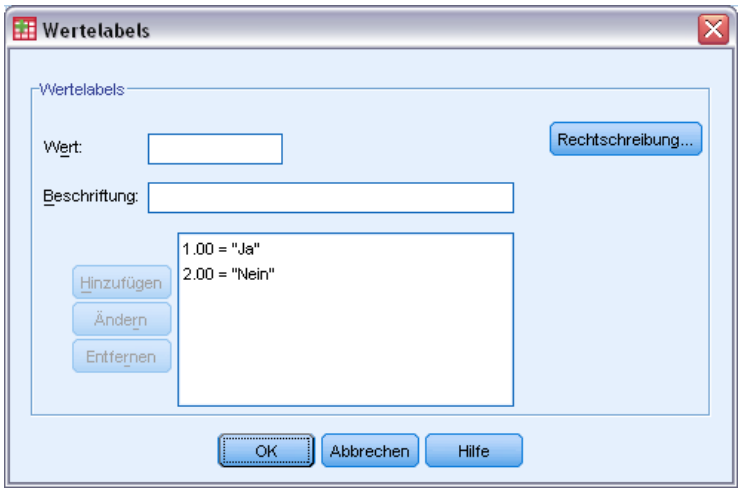
Die Benutzeroberfläche des Dialogfelds “Entscheidungsbaum” geht davon aus, dass entweder für *alle* nichtfehlenden Werte einer kategorialen (nominalen, ordinalen) abhängigen Variablen Wertelabels definiert sind oder für *keine*. Einige Funktionen sind nicht verfügbar, wenn nicht mindestens zwei nichtfehlende Werte der kategorialen abhängigen Variablen Wertelabels aufweisen. Wenn für mindestens zwei nichtfehlende Werte Wertelabels definiert sind, werden alle Fälle mit anderen Werten, die keine Wertelabels aufweisen, aus der Analyse ausgeschlossen.

Die ursprüngliche Datendatei in diesem Beispiel enthält keine definierten Wertelabels und wenn die abhängige Variable als nominal behandelt wird, verwendet das Baummodell alle nichtfehlenden Werte in der Analyse. In diesem Beispiel sind diese Werte 1, 2 und 3.

Was geschieht aber, wenn wir Wertelabels für einige, jedoch nicht für alle, Werte der abhängigen Variablen definieren?

- Klicken Sie im Fenster “Daten-Editor” auf die Registerkarte Variablenansicht.
- Klicken Sie auf die Zelle Werte für die Variable *abhängig*.

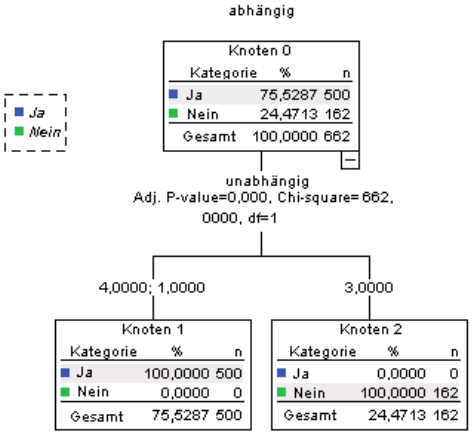
Abbildung 3-5
Definieren von Wertelabels für die Variable “dependent” (abhängig)



- Geben Sie zunächst 1 als Wert und Ja als Wertelabel ein und klicken Sie dann auf Hinzufügen.
- Geben Sie danach 2 als Wert und Nein als Wertelabel ein und klicken Sie dann auf Hinzufügen.
- Klicken Sie anschließend auf OK.

- Öffnen Sie noch einmal das Dialogfeld “Entscheidungsbaum”. Im Dialogfeld sollte noch immer *abhängig* als abhängige Variable mit nominalem Messniveau ausgewählt sein.
- Klicken Sie auf OK, um die Prozedur noch einmal auszuführen.

Abbildung 3-6
Baum für nominale abhängige Variable, teilweise mit Wertelabels



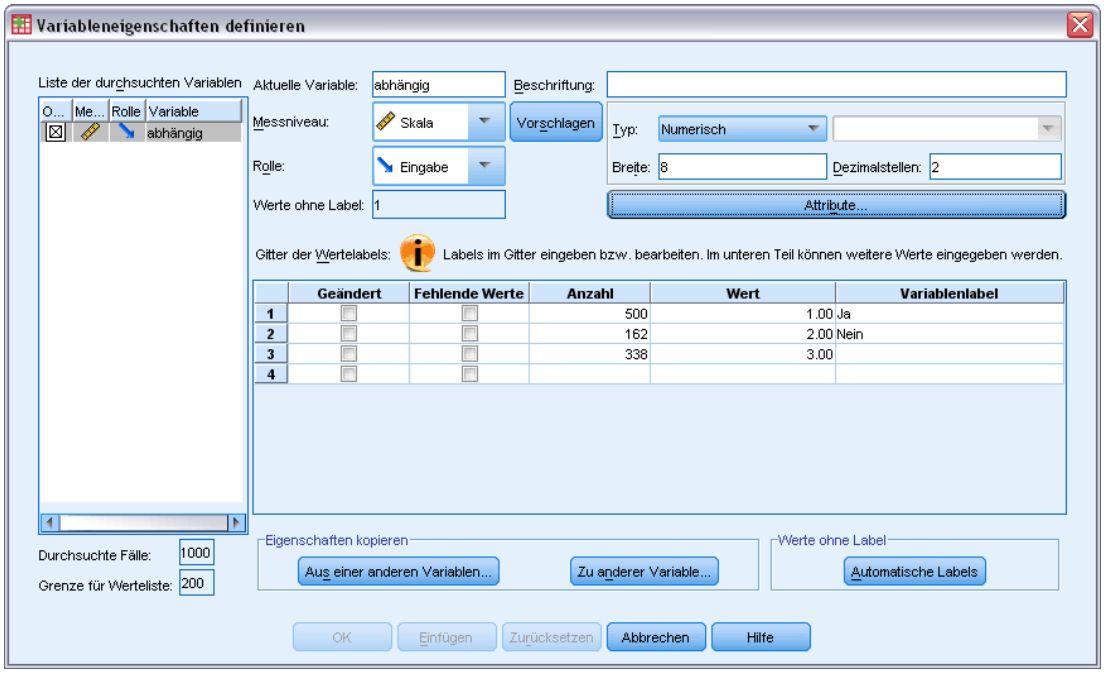
Nun sind nur die beiden abhängigen Variablenwerte mit definierten Variablenlabels im Baummodell enthalten. Alle Fälle mit dem Wert 3 für die abhängige Variable wurden ausgeschlossen, was Ihnen möglicherweise nicht sofort auffällt, wenn Sie mit den Daten nicht vertraut sind.

Zuweisen von Wertelabels zu allen Werten

Um einen versehentlichen Ausschluss gültiger kategorialer Werte aus der Analyse zu vermeiden, müssen Sie mit der Option “Variableneigenschaften definieren” allen abhängen Variablenwerten, die in den Daten gefunden werten, Wertelabel zuordnen.

Wenn die Informationen aus dem Datenlexikon für die Variable *name* im Dialogfeld "Variableneigenschaften definieren" angezeigt werden, können Sie sehen, dass es zwar mehr als 300 Fälle mit den Wert 3 für diese Variable gibt, jedoch kein Wertelabel für diesen Wert definiert wurde.

Abbildung 3-7
Variable, teilweise mit Wertelabels, im Dialogfeld "Variableneigenschaften definieren"



Kapitel
4

Verwenden von Entscheidungsbäumen zur Bewertung des Kreditrisikos

Eine Bank unterhält eine Datenbank mit Informationen zu Kunden, die Kredite von der Bank aufgenommen haben, einschließlich der Informationen, ob sie die Kredite zurückgezahlt haben oder ihren Zahlungsverpflichtungen nicht nachgekommen sind. Mithilfe von Entscheidungsbäumen können Sie die Merkmale der beiden Kundengruppen analysieren und Modelle konstruieren, mit denen sich die Wahrscheinlichkeit voraussagen lässt, dass Kreditantragsteller ihre Kredite nicht zurückzahlen.

Die Kreditdaten sind in der Datei *tree_credit.sav* gespeichert. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Decision Trees (Entscheidungsbäume) 18*.

Erstellen des Modells

Die Prozedur “Entscheidungsbaum” bietet mehrere verschiedene Methoden zur Erstellung von Baummodellen. In diesem Beispiel verwenden wir die Standardmethode:

CHAID. Steht für "Chi-squared Automatic Interaction Detection", d. h. automatische Entdeckung von Zusammenhängen mittels Chi-Quadrat-Tests. In jedem Schritt bestimmt das CHAID-Verfahren diejenige unabhängige Variable (Einflussvariable/Prädiktor), die den stärksten Zusammenhang mit der abhängigen Variablen aufweist. Die Kategorien der einzelnen Einflussvariablen werden zusammengeführt, wenn sie im Hinblick auf die abhängige Variable nicht signifikant unterschiedlich sind.

Erstellen des CHAID-Baummodells

- Zum Erstellen einer Entscheidungsbaum-Analyse wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Klassifizieren
    Baum...

Abbildung 4-1
Dialogfeld "Entscheidungsbaum"

- Wählen Sie *Kreditrating* als abhängige Variable aus.
- Wählen Sie alle verbleibenden Variablen als unabhängige Variablen aus. (Die Prozedur schließt automatisch alle Variablen aus, die keinen signifikanten Beitrag zum endgültigen Modell leisten.)

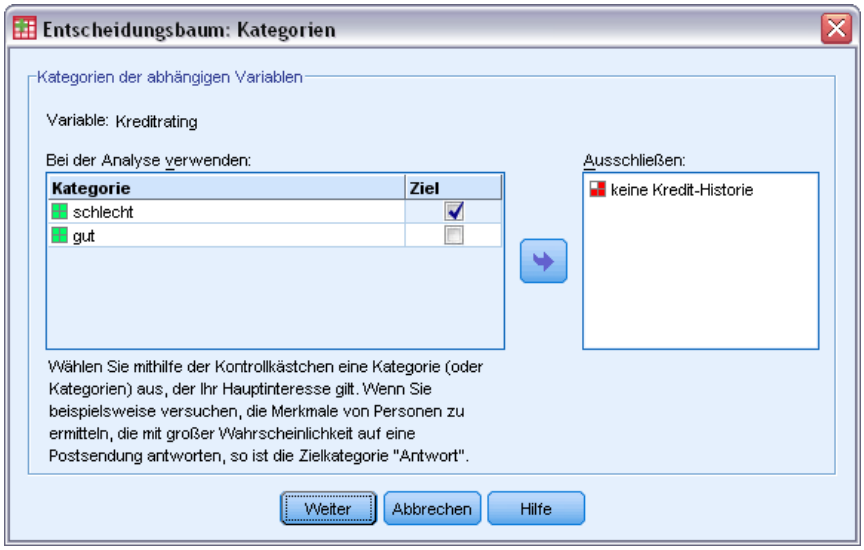
  Zu diesem Zeitpunkt könnten Sie die Prozedur ausführen und ein grundlegendes Baummodell erstellen, doch wir wählen weitere Ausgaben aus und nehmen einige kleinere Anpassungen an den Kriterien vor, die für die Erstellung des Modells verwendet wurden.

Auswahl der Zielkategorien

- Klicken Sie auf die Schaltfläche Kategorien unmittelbar unterhalb der ausgewählten abhängigen Variablen.

Dadurch wird das Dialogfeld “Kategorien” geöffnet, in dem Sie die relevanten Zielkategorien der abhängigen Variablen angeben können. Zielkategorien betreffen nicht das Baummodell selbst, sondern bestimmte Ausgaben, und Optionen sind nur verfügbar, wenn Zielkategorien ausgewählt wurden.

Abbildung 4-2
Dialogfeld “Kategorien”


Entscheidungsbaum: Kategorien
Kategorien der abhängigen Variablen
Variable: Kreditrating
Bei der Analyse verwenden:
Kategorie
Ziel
schlecht
gut
Ausschließen:
keine Kredit-Historie
Wählen Sie mithilfe der Kontrollkästchen eine Kategorie (oder Kategorien) aus, der Ihr Hauptinteresse gilt. Wenn Sie beispielsweise versuchen, die Merkmale von Personen zu ermitteln, die mit großer Wahrscheinlichkeit auf eine Postsendung antworten, so ist die Zielkategorie "Antwort".
Weiter
Abbrechen
Hilfe

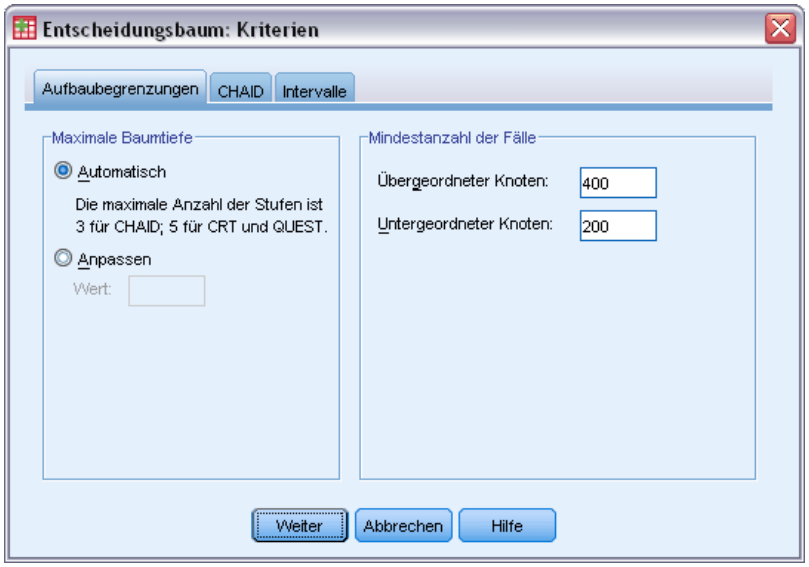
- Aktivieren Sie das Kontrollkästchen “Ziel” für die Kategorie *Schlecht*. Kunden mit schlechtem Kreditrating (ein Kredit wurde nicht zurückgezahlt) werden als relevante Zielkategorie behandelt.
- Klicken Sie auf Weiter.

Angeben von Aufbaukriterien für Bäume

In diesem Beispiel möchten wir den Baum ziemlich einfach halten. Daher begrenzen wir den Aufbau des Baums durch Anhebung der Mindestanzahl der Fälle für über- und untergeordnete Knoten.

- Klicken Sie im Hauptdialogfeld “Entscheidungsbaum” auf Kriterien.

Abbildung 4-3
Dialogfeld "Kriterien", Registerkarte "Aufbaubegrenzungen"



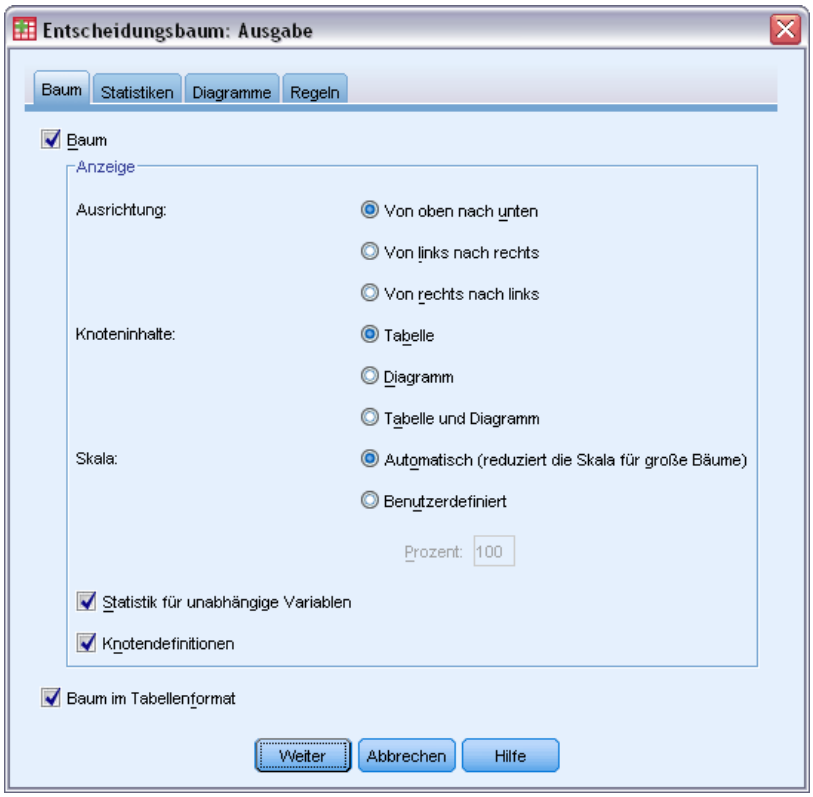
- Geben Sie im Gruppenfeld "Mindestanzahl der Fälle" den Wert 400 für den übergeordneten und den Wert 200 für den untergeordneten Knoten ein.
- Klicken Sie auf Weiter.

Auswahl zusätzlicher Ausgaben

- Klicken Sie im Dialogfeld "Entscheidungsbaum" auf Ausgabe.

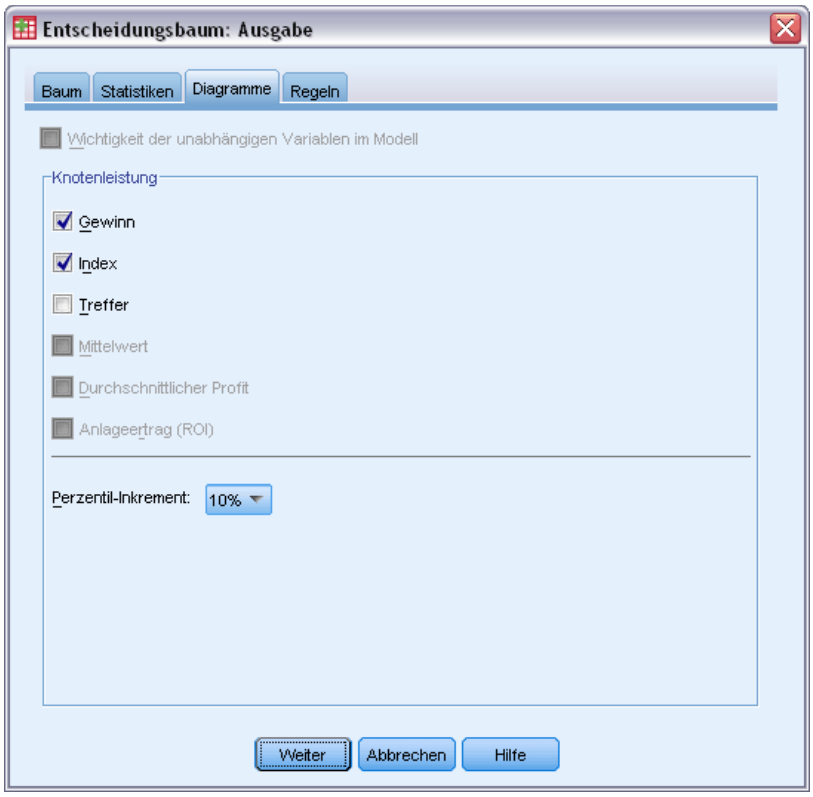
Dadurch wird ein Dialogfeld mit Registerkarten geöffnet, in dem verschiedene Typen von zusätzlichen Ausgaben ausgewählt werden können.

Abbildung 4-4
Dialogfeld "Ausgabe", Registerkarte "Baum"



- Aktivieren Sie auf der Registerkarte "Baum" die Option Baum im Tabellenformat.
- Klicken Sie anschließend auf die Schaltfläche Diagramme.

Abbildung 4-5
Dialogfeld "Ausgabe", Registerkarte "Diagramme"



- Aktivieren Sie Gewinn und Index.

  Hinweis: Bei diesen Diagrammen ist eine Zielkategorie für die abhängige Variable erforderlich. In diesem Beispiel kann die Registerkarte "Diagramme" erst aufgerufen werden, nachdem Sie mindestens eine Zielkategorie angegeben haben.

- Klicken Sie auf Weiter.

Speichern vorhergesagter Werte

Sie können Variablen speichern, die Informationen über Modellvorhersagen enthalten. Sie können beispielsweise das für die einzelnen Fälle vorhergesagte Kreditrating speichern und anschließend diese Vorhersagen mit dem tatsächlichen Kreditrating vergleichen.

- Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf Speichern.

Abbildung 4-6
Dialogfeld "Speichern"

- Wählen Sie die Optionen Endknotennummer, Vorhergesagter Wert und Vorhergesagte Wahrscheinlichkeiten aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Entscheidungsbaum" auf OK, um die Prozedur auszuführen.

Bewertung des Modells

In diesem Beispiel beinhalten die Modellergebnisse folgende Elemente:

- Tabellen mit Informationen über das Modell
- Baumdiagramm
- Grafiken, die die Leistungsfähigkeit des Modells anzeigen
- In die Arbeitsdatei aufgenommene Modellvorhersagevariablen

Modellzusammenfassungstabelle

Abbildung 4-7
Modellzusammenfassung

Spezifikationen	Aufbaumethode	CHAID
	Abhängige Variable	Kreditrating
	Unabhängige Variablen	Alter, Einkommen in Kategorien, Anzahl an Kreditkarten, Ausbildung, Autodarlehen
	Validierung	NONE
	Maximale Baumtiefe	3
	Mindestanzahl der Fälle im übergeordneten Knoten	400
	Mindestanzahl der Fälle im untergeordneten Knoten	200
Ergebnisse	Aufgenommene unabhängige Variablen	Einkommen in Kategorien, Anzahl an Kreditkarten, Alter
	Anzahl der Knoten	10
	Anzahl der Endknoten	6
	Tiefe	3

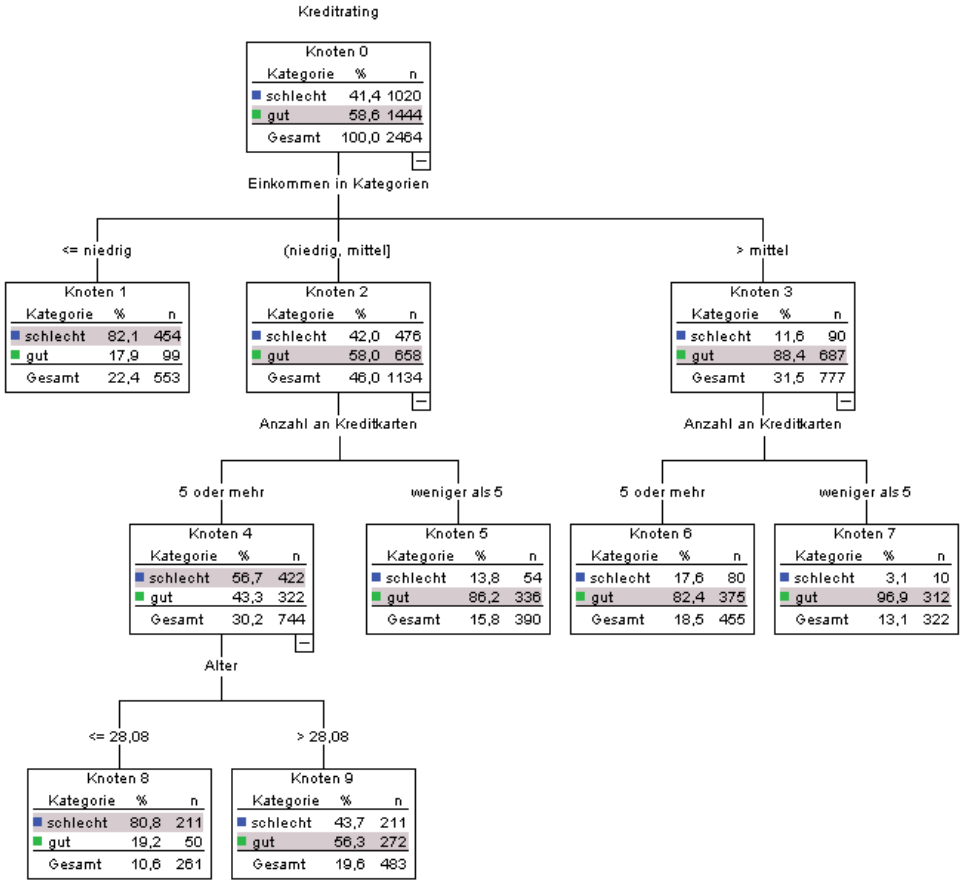
Die Modellzusammenfassungstabelle bietet sehr allgemeine Informationen über die für die Konstruktion des Modells verwendeten Spezifikationen und das resultierende Modell.

- Der Abschnitt “Spezifikationen” bietet Informationen zu den bei der Konstruktion des Baummodells verwendeten Einstellungen, einschließlich der bei der Analyse verwendeten Variablen.
- Der Abschnitt “Ergebnisse” bietet Informationen zur Gesamtanzahl der Knoten und zur Anzahl der Endknoten, zur Tiefe des Baums (Anzahl der Ebenen unterhalb des Stammknotens) und zu den im endgültigen Modell enthaltenen unabhängigen Variablen.

Es wurden fünf unabhängige Variablen angegeben, jedoch nur drei wurden in das endgültige Modell aufgenommen. Die Variablen für *Ausbildung* und Anzahl der laufenden *Autodarlehen* leisteten keinen signifikanten Beitrag zum Modell und wurden daher beim endgültigen Modell automatisch weggelassen.

Baumdiagramm

Abbildung 4-8
Baumdiagramm für die Erstellung eines Modells für das Kreditrating



Das Baumdiagramm ist eine grafische Darstellung des Baummodells. Dieses Baumdiagramm zeigt Folgendes:

- Bei Verwendung der CHAID-Methode ist *Einkommen in Kategorien* die beste Einflussvariable für *Kreditrating*.
- Bei der unteren Einkommensklasse ist *Einkommen in Kategorien* die einzige signifikante Einflussvariable für *Kreditrating*. Von den Bankkunden in dieser Kategorie haben 82 % Kredite nicht zurückgezahlt. Da unterhalb dieses Knotens keine untergeordneten Knoten vorhanden sind, wird dieser Knoten als **End**knoten betrachtet.
- Bei der mittleren und der hohen Einkommensklasse ist die nächstbeste Einflussvariable *Anzahl an Kreditkarten*.
- Bei Kunden mit mittlerem Einkommen und mindestens fünf Kreditkarten enthält das Modell eine weitere Einflussvariable: *Alter*. Über 80 % dieser Kunden, die 28 Jahre oder jünger waren, hatten ein schlechtes Kreditrating, wohingegen nur knapp die Hälfte der Kunden über 28 aus dieser Gruppe ein schlechtes Kreditrating aufwiesen.

Mit dem Baumeditor können Sie ausgewählte Zweige aus- und einblenden, Farben und Schriftarten ändern und Untergruppen von Fällen auf der Grundlage der ausgewählten Knoten auswählen. Für weitere Informationen siehe Auswählen der Fälle in Knoten auf S. 74.

Baumtabelle

Abbildung 4-9
Baumtabelle für das Kreditrating

	schlecht		gut		Gesamt		Vorhergesagte Kategorie	Übergeordneter Knoten
Knoten	N	Prozent	N	Prozent	N	Prozent		
0	1020	41,4%	1444	58,6%	2464	100,0%	gut	
1	454	82,1%	99	17,9%	553	22,4%	schlecht	0
2	476	42,0%	658	58,0%	1134	46,0%	gut	0
3	90	11,6%	687	88,4%	777	31,5%	gut	0
4	422	56,7%	322	43,3%	744	30,2%	schlecht	2
5	54	13,8%	336	86,2%	390	15,8%	gut	2
6	80	17,6%	375	82,4%	455	18,5%	gut	3
7	10	3,1%	312	96,9%	322	13,1%	gut	3
8	211	80,8%	50	19,2%	261	10,6%	schlecht	4
9	211	43,7%	272	56,3%	483	19,6%	gut	4

Die Baumtabelle bietet, wie der Name schon sagt, die wichtigsten Informationen aus dem Baumdiagramm in Tabellenform. Für jeden Knoten wird in der Tabelle Folgendes angezeigt:

- Die Anzahl und der Prozentsatz der Fälle in jeder Kategorie der abhängigen Variablen.
- Die vorhergesagte Kategorie für die abhängige Variable. In diesem Beispiel handelt es sich bei der vorhergesagten Kategorie um die Kategorie *Kreditrating* mit mehr als 50 % der Fälle in diesem Knoten, da es nur zwei mögliche Kreditratings gibt.
- Der übergeordnete Knoten für jeden Knoten im Baum. Beachten Sie, dass Knoten 1 – der Knoten für das niedrige Einkommensniveau – für keinen anderen Knoten als übergeordneter Knoten fungiert. Da es sich um einen Endknoten handelt, besitzt er keine untergeordneten Knoten.

Abbildung 4-10
Baumtabelle für das Kreditrating (Fortsetzung)

Primäre unabhängige Variable				
Variable	Sig.[a]	Chi-Quadrat	df	Aufteilungwerte
Einkommen in Kategorien	,000	662,457	2	<= niedrig
Einkommen in Kategorien	,000	662,457	2	(niedrig, mittel]
Einkommen in Kategorien	,000	662,457	2	> mittel
Anzahl an Kreditkarten	,000	193,113	1	5 oder mehr
Anzahl an Kreditkarten	,000	193,113	1	weniger als 5
Anzahl an Kreditkarten	,000	38,587	1	5 oder mehr
Anzahl an Kreditkarten	,000	38,587	1	weniger als 5
Alter	,000	95,299	1	<= 28,0792
Alter	,000	95,299	1	> 28,0792

- Die unabhängige Variable, die zur Aufteilung des Knotens verwendet wird.

- Der Chi-Quadrat-Wert (da der Baum mit der Methode "CHAID" erstellt wurde), die Freiheitsgrade (*df*) und das Signifikanzniveau (*Sig.*) für die Aufteilung. Für die meisten Zwecke sind Sie vermutlich nur am Signifikanzniveau interessiert, das für alle Aufteilungen weniger als 0,0001 beträgt.
- Die Werte der unabhängigen Variablen für diesen Knoten.

Hinweis: Bei unabhängigen ordinalen und metrischen Variablen können im Baum und in der Baumtabelle Bereiche in der allgemeinen Form *(Wert1, Wert2]* ausgedrückt werden, die bedeutet: "größer als Wert1 und kleiner oder gleich Wert2". In diesem Beispiel gibt es für das Einkommensniveau nur drei mögliche Werte – *Niedrig*, *Mittel* und *Hoch* – und *(Low, Medium]* ((Niedrig, Mittel]) bedeutet einfach *Medium* (Mittel). *>Mittel* bedeutet *Hoch*.

Gewinne für Knoten

Abbildung 4-11
Gewinne für Knoten

	Knoten		Gewinn			
Knoten	N	Prozent	N	Prozent	Treffer	Index
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Die Tabelle "Gewinne für Knoten" bietet eine Zusammenfassung der Informationen über die Endknoten im Modell.

- Nur die Endknoten – Knoten, an denen der Baum nicht mehr weiter wächst – werden in der Tabelle aufgeführt. In den meisten Fällen sind nur die Endknoten von Interesse, da sie die besten Klassifikationsvoraussagen für das Modell darstellen.
- Da die Gewinnwerte Informationen zu Zielkategorien bieten, ist diese Tabelle nur verfügbar, wenn mindestens eine Zielkategorie angegeben wurde. In diesem Beispiel gibt es nur eine einzige Zielkategorie und damit nur eine einzige Tabelle für die Gewinne für die Knoten.
- *Knoten: N* ist die Anzahl der Fälle in den einzelnen Endknoten und *Knoten: Prozent* ist der Prozentsatz der Gesamtzahl der Fälle in den einzelnen Knoten.
- *Gewinn: N* ist die Anzahl der Fälle in jedem Endknoten in der Zielkategorie und *Gewinn: Prozent* ist der Prozentsatz der Fälle in der Zielkategorie bezogen auf die Gesamtzahl der Fälle in der Zielkategorie – in diesem Beispiel die Anzahl und der Prozentsatz der Fälle mit schlechtem Kreditrating.
- Bei kategorialen abhängigen Variablen ist *Antwort* der Prozentsatz der Fälle im Knoten der angegebenen Zielkategorie. In diesem Beispiel handelt es sich hierbei um dieselben Prozentsätze, die im Baumdiagramm für die Kategorie *Schlecht* angezeigt wurden.
- Bei kategorialen abhängigen Variablen ist *Index* das Verhältnis des Antwortprozentsatzes für die Zielkategorie im Vergleich zum Antwortprozentsatz für die gesamte Stichprobe.

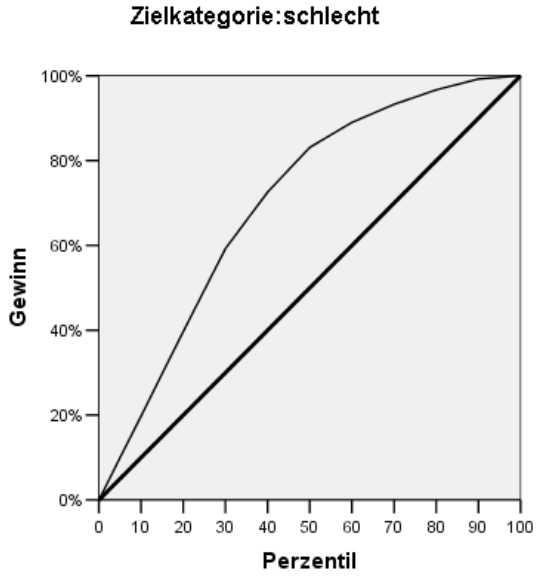
Indexwerte

Der Indexwert zeigt an, wie weit der *beobachtete* Prozentsatz für die Zielkategorie bei diesem Knoten von dem *erwarteten* Prozentsatz für die Zielkategorie abweicht. Der Prozentsatz für die Zielkategorie im Stammknoten steht für den erwarteten Prozentsatz vor der Berücksichtigung der Effekte der unabhängigen Variablen.

Ein Indexwert von mehr als 100 % bedeutet, dass die Zielkategorie mehr Fälle aufweist als den Gesamtprozentsatz in der Zielkategorie. Umgekehrt bedeutet ein Indexwert von weniger als 100 %, dass sich in der Zielkategorie weniger Fälle befinden als der Gesamtprozentsatz.

Gewinndiagramm

Abbildung 4-12
Gewinndiagramm für die Zielkategorie für schlechtes Kreditrating

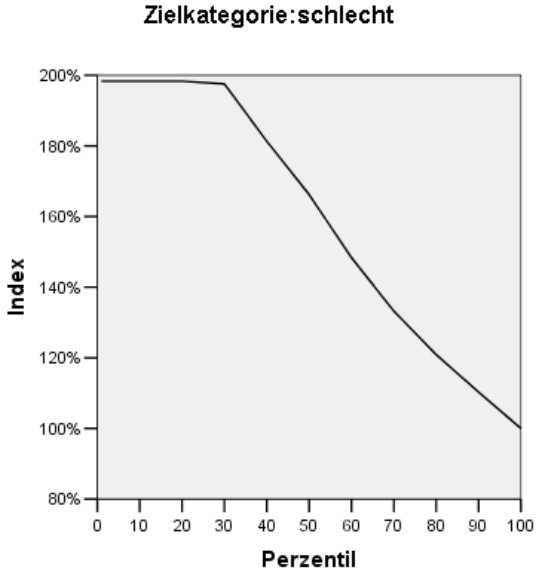


Dieses Gewinndiagramm zeigt an, dass das Modell ziemlich gut ist.

Kumulative Gewinndiagramme beginnen immer bei 0 % und enden bei 100 %. Bei einem guten Modell steigt die Gewinntabelle steil in Richtung 100 % an und flacht dann ab. Ein Modell, das keine Informationen bietet, folgt der diagonalen Bezugslinie.

Indexdiagramm

Abbildung 4-13
Indexdiagramm für die Zielkategorie für schlechtes Kreditrating



Das Indexdiagramm zeigt ebenfalls an, dass das Modell gut ist. Kumulative Indexdiagramme starten in der Regel bei über 100 % und fallen langsam bis auf 100 % ab.

Bei einem guten Modell sollte der Indexwert deutlich oberhalb von 100 % beginnen, eine Weile auf hohem Niveau bleiben und dann steil auf 100 % absinken. Bei einem Modell, das keine Informationen bietet, bleibt die Linie im gesamten Diagramm bei ca. 100 %.

Risikoschätzer und Klassifizierung

Abbildung 4-14
Tabellen für Risiko und Klassifizierung

Risiko

Schätzer	Standardfehler
,205	,008

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Klassifikation

	Vorhergesagt		
Beobachtet	schlecht	gut	Prozent korrekt
schlecht	665	355	65,2%
gut	149	1295	89,7%
Gesamtprozentsatz	33,0%	67,0%	79,5%

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

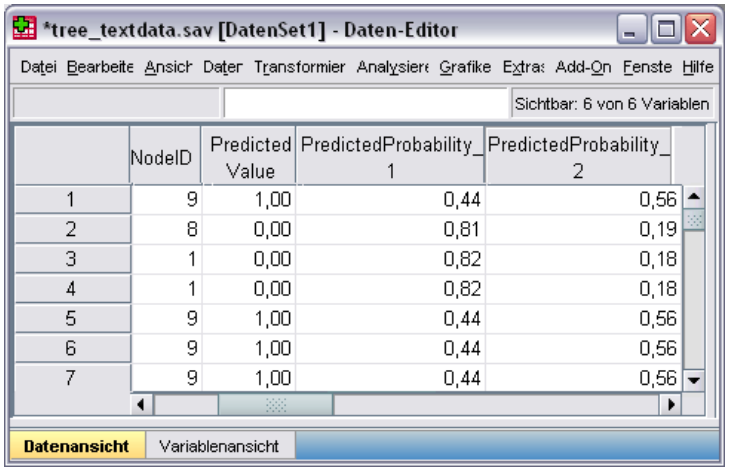
Die Tabellen für Risiko und Klassifizierung ermöglichen eine schnelle Einschätzung der Güte des Modells.

- Der Risikoschätzer 0,205 zeigt an, dass die vom Modell vorhergesagte Kategorie (gutes oder schlechtes Kreditrating) in 20,5 % der Fälle falsch ist. Das Risiko der Fehlklassifizierung eines Kunden liegt also bei etwa 21 %.
- Die Ergebnisse in der Klassifikationstabelle sind mit dem Risikoschätzer konsistent. Die Tabelle zeigt, dass das Modell ca. 79,5 % der Kunden richtig klassifiziert.

Die Klassifikationstabelle zeigt jedoch ein potenzielles Problem bei diesem Modell: bei den Kunden mit schlechtem Kreditrating sagt es nur für 65 % eine schlechte Bewertung voraus, was bedeutet, dass 35 % der Kunden mit schlechtem Kreditrating fälschlicherweise bei den “guten” Kunden eingeordnet werden.

Vorhergesagte Werte

Abbildung 4-15
Neue Variablen für vorhergesagte Werte und Wahrscheinlichkeiten


*tree_textdata.sav [DatenSet1] - Daten-Editor
Datei Bearbeite Ansicht Daten Transformier Analysier Grafike Extras Add-On Fenste Hilfe
Sichtbar: 6 von 6 Variablen
Datenansicht
Variablenansicht

	NodeID	Predicted Value	PredictedProbability_1	PredictedProbability_2
1	9	1,00	0,44	0,56
2	8	0,00	0,81	0,19
3	1	0,00	0,82	0,18
4	1	0,00	0,82	0,18
5	9	1,00	0,44	0,56
6	9	1,00	0,44	0,56
7	9	1,00	0,44	0,56

In der Arbeitsdatei wurden vier neue Variablen erstellt:

NodeID. Die Nummer des Endknotens für jeden Fall.

PredictedValue. Der vorhergesagte Wert der abhängigen Variablen für jeden Fall. Da die abhängige Variable als 0 = *Schlecht* und 1 = *Gut* kodiert ist, bedeutet ein vorhergesagter Wert 0, dass für den Fall ein schlechtes Kreditrating vorhergesagt wird.

PredictedProbability. Die Wahrscheinlichkeit, dass der Fall in die einzelnen Kategorien der abhängigen Variablen gehört. Da es nur zwei mögliche Werte für die abhängige Variable gibt, werden zwei Variablen erstellt:

- **PredictedProbability_1.** Die Wahrscheinlichkeit, dass der Fall in die Kategorie für schlechtes Kreditrating gehört.
- **PredictedProbability_2.** Die Wahrscheinlichkeit, dass der Fall in die Kategorie für gutes Kreditrating gehört.

Die vorhergesagte Wahrscheinlichkeit ist einfach der Anteil der Fälle in den einzelnen Kategorien der abhängigen Variablen für den Endknoten, der den jeweiligen Fall enthält. In Knoten 1 beispielsweise befinden sich 82 % der Fälle in der schlechten Kategorie und 18 % der Fälle in der guten Kategorie, was eine vorhergesagte Wahrscheinlichkeit von 0,82 bzw. 0,18 ergibt.

Bei einer kategorialen abhängigen Variablen ist der vorhergesagte Wert die Kategorie mit dem höchsten Anteil von Fällen im Endknoten für den jeweiligen Fall. Beispiel: Beim ersten Fall ist der vorhergesagte Wert 1 (gutes Kreditrating), da ca. 56 % der Fälle in seinem Endknoten ein gutes Kreditrating aufweisen. Umgekehrt ist beim zweiten Fall der vorhergesagte Wert 0 (schlechtes Kreditrating), da ca. 81 % der Fälle in seinem Endknoten ein schlechtesKreditrating aufweisen.

Wenn Sie jedoch Kosten definiert haben, ist die Beziehung zwischen vorhergesagter Kategorie und vorhergesagten Wahrscheinlichkeiten möglicherweise nicht so offensichtlich. Für weitere Informationen siehe Zuweisen von Kosten zu den Ergebnissen auf S. 78.

Verfeinern des Modells

Insgesamt weist das Modell eine Quote für die korrekte Klassifizierung von knapp unter 80 % auf. Dies spiegelt sich in den meisten Endknoten wider, in denen die vorhergesagte Kategorie – die markierte Kategorie im Knoten – in mindestens 80 % der Fälle mit der tatsächlichen Kategorie übereinstimmt.

Es gibt jedoch einen Endknoten, in dem die Fälle ziemlich gleichmäßig zwischen gutem und schlechtem Kreditrating verteilt sind. In Knoten 9 ist das vorhergesagtes Kreditrating “gut”, doch nur 56 % der Fälle in diesem Knoten weisen tatsächlich ein gutes Kreditrating auf. Das bedeutet, dass fast die Hälfte der Fälle in diesem Knoten (44 %) die falsche vorhergesagte Kategorie aufweisen. Wenn das Hauptziel darin besteht, das Risiko für schlechtes Kreditrating zu ermitteln, leistet dieser Knoten keinen guten Beitrag.

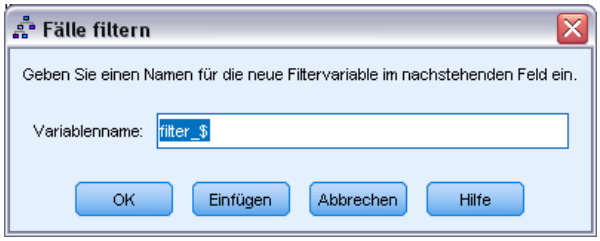
Auswählen der Fälle in Knoten

Betrachten wir die Fälle in Knoten 9, um zu ermitteln, ob die Daten irgendwelche zusätzlichen Informationen bieten, die von Nutzen sein könnten.

- Doppelklicken Sie auf den Baum im Viewer, um den Baum-Editor zu öffnen.
- Klicken Sie auf Knoten 9, um ihn auszuwählen. (Mehrere Knoten können Sie auswählen, indem Sie beim Klicken die STRG-Taste gedrückt halten.)
- Wählen Sie die folgenden Menübefehle des Baum-Editors aus:

Regeln
  Fälle filtern...

Abbildung 4-16
Dialogfeld "Fälle filtern"



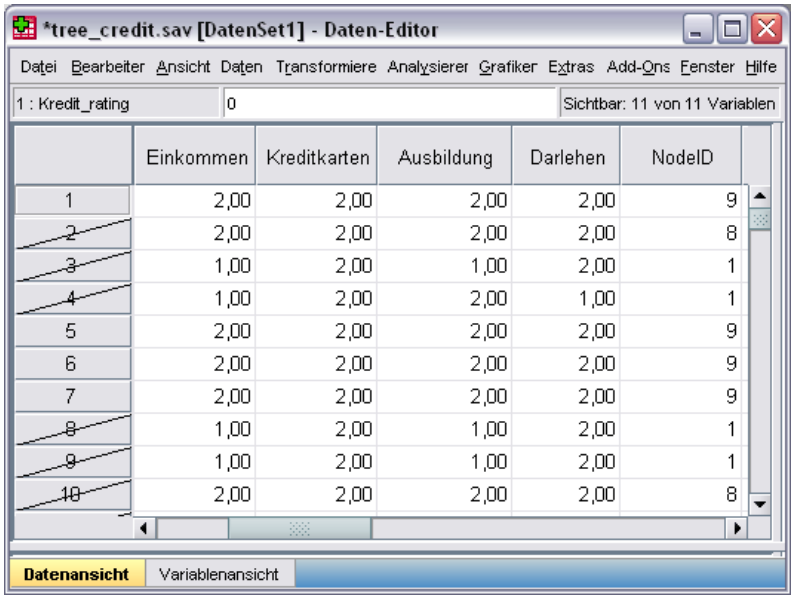
Das Dialogfeld "Fälle filtern" erstellt eine Filtervariable und wendet eine Filtereinstellung auf der Grundlage der Werte der betreffenden Variablen an. Standardmäßig lautet der Name der Filtervariablen *filter_$*.

- Die Fälle aus den ausgewählten Knoten erhalten für die Filtervariable den Wert 1.
- Alle anderen Fälle erhalten den Wert 0 und werden aus den nachfolgenden Analysen ausgeschlossen, bis Sie den Filterstatus ändern.

In diesem Beispiel bedeutet dies, dass die Fälle, die sich nicht in Knoten 9 befinden, vorerst herausgefiltert (jedoch nicht gelöscht) werden.

▶ Klicken Sie auf OK, um die Filtervariable zu erstellen und die Filterbedingung anzuwenden.

Abbildung 4-17
Gefilterte Fälle im Daten-Editor


*tree_credit.sav [DatenSet1] - Daten-Editor
Datei Bearbeiter Ansicht Daten Transformiere Analysieren Grafiker Extras Add-Ons Fenster Hilfe
1 : Kredit_rating
0
Sichtbar: 11 von 11 Variablen
Einkommen Kreditkarten Ausbildung Darlehen NodeID
1 2,00 2,00 2,00 2,00 9
2 2,00 2,00 2,00 2,00 8
3 1,00 2,00 1,00 2,00 1
4 1,00 2,00 2,00 1,00 1
5 2,00 2,00 2,00 2,00 9
6 2,00 2,00 2,00 2,00 9
7 2,00 2,00 2,00 2,00 9
8 1,00 2,00 1,00 2,00 1
9 1,00 2,00 1,00 2,00 1
10 2,00 2,00 2,00 2,00 8
Datenansicht
Variablenansicht

Im Daten-Editor werden Fälle, die herausgefiltert wurden, mit einem diagonalen Strich durch die Zeilennummer gekennzeichnet. Die Fälle, die sich nicht in Knoten 9 befinden, werden herausgefiltert. Die Fälle in Knoten 9 werden nicht gefiltert; daher enthalten alle nachfolgenden Analysen nur Fälle aus Knoten 9.

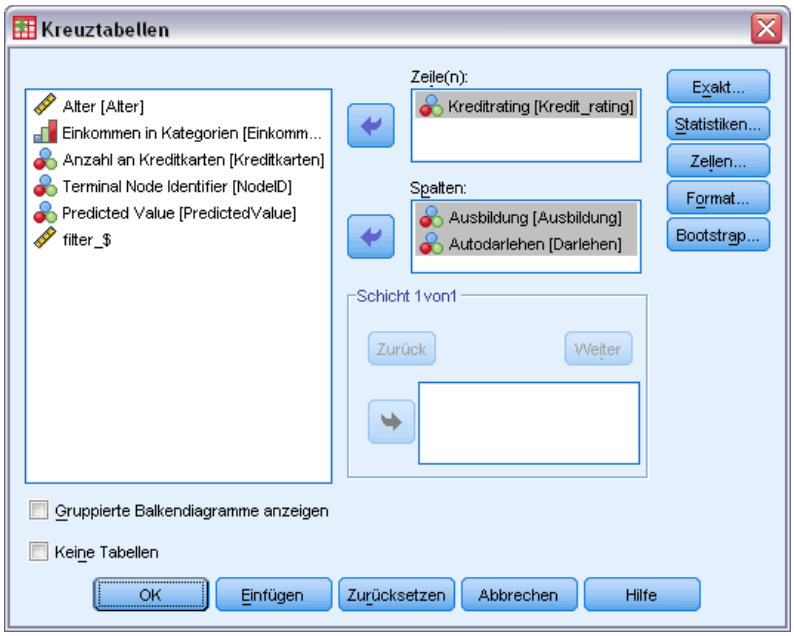
Untersuchung der ausgewählten Fälle

Als ersten Schritt bei der Untersuchung der Fälle in Knoten 9 sollten Sie die Variablen betrachten, die nicht im Modell verwendet wurden. In diesem Beispiel wurden alle Variablen in der Datendatei in die Analyse aufgenommen, zwei davon wurden jedoch nicht in das endgültige Modell aufgenommen: *Ausbildung* und *Autodarlehen*. Da es vermutlich einen guten Grund dafür gab, dass die Prozedur sie beim endgültigen Modell nicht verwendete, sind sie vermutlich nicht sonderlich aussagekräftig. Wir wollen sie uns jedoch dennoch einmal genauer anschauen.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Deskriptive Statistiken
    Kreuztabellen...

Abbildung 4-18
Dialogfeld "Kreuztabellen"

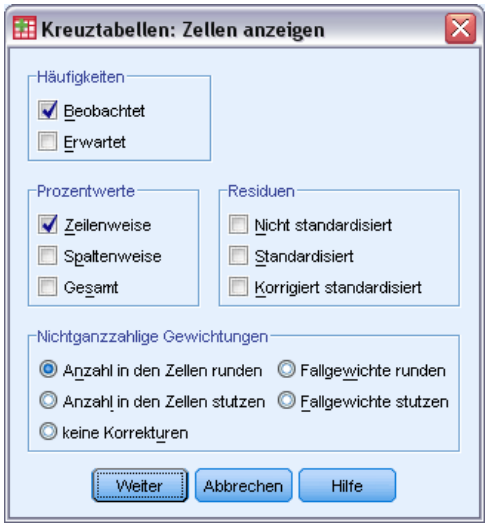


▶ Wählen Sie *Kreditrating* als Zeilenvariable aus.

▶ Wählen Sie *Ausbildung* und *Autodarlehen* als Spaltenvariablen aus.

▶ Klicken Sie auf Zellen.

Abbildung 4-19
Dialogfeld "Kreuztabellen: Zellenanzeige"



- Aktivieren Sie im Gruppenfeld "Prozentwerte" die Option Zeilenweise.
- Klicken Sie anschließend auf Weiter und danach im Hauptdialogfeld von "Kreuztabellen" auf OK, um die Prozedur auszuführen.

Bei der Untersuchung der Kreuztabellen wird ersichtlich, dass für die beiden nicht im Modell enthaltenen Variablen kein großer Unterschied zwischen den Fällen in den Kategorien für gutes und schlechtes Kreditrating besteht.

Abbildung 4-20
Kreuztabellen für die Fälle im ausgewählten Knoten

Kreditrating * Ausbildung Kreuztabelle

			Ausbildung		
			Schulabschluß	Universitätsabschluß	Gesamt
Kreditrating	schlecht	Anzahl	513	507	1020
		% von Kreditrating	50,3%	49,7%	100,0%
	gut	Anzahl	717	727	1444
		% von Kreditrating	49,7%	50,3%	100,0%
Gesamt		Anzahl	1230	1234	2464
		% von Kreditrating	49,9%	50,1%	100,0%

Kreditrating * Autodarlehen Kreuztabelle

			Autodarlehen		
			0 oder 1	2 oder mehr	Gesamt
Kreditrating	schlecht	Anzahl	178	842	1020
		% von Kreditrating	17,5%	82,5%	100,0%
	gut	Anzahl	715	729	1444
		% von Kreditrating	49,5%	50,5%	100,0%
Gesamt		Anzahl	893	1571	2464
		% von Kreditrating	36,2%	63,8%	100,0%

- Was *Ausbildung* betrifft, so besitzt etwas mehr als die Hälfte der Fälle mit schlechtem Kreditrating nur einen Schulabschluss, während etwas mehr als die Hälfte mit gutem Kreditrating einen Universitätsabschluss vorzuweisen hat, doch dieser Unterschied ist nicht statistisch signifikant.
- Was *Autodarlehen* betrifft, so ist der Prozentsatz der Fälle mit gutem Kreditrating, die höchstens ein einziges Autodarlehen haben, höher als der entsprechende Prozentsatz für die Fälle mit schlechtem Kreditrating, doch die überwältigende Mehrheit in beiden Gruppen hat mindestens zwei Autodarlehen.

Sie können nun zwar besser nachvollziehen, warum diese Variablen nicht in das endgültige Modell aufgenommen wurden, es ist jedoch leider nicht klarer geworden, wie eine bessere Vorhersage für Knoten 9 erzielt werden könnte. Wenn es andere Variablen gäbe, die nicht für die Analyse spezifiziert wurden, sollten Sie diese eventuell untersuchen, bevor Sie fortfahren.

Zuweisen von Kosten zu den Ergebnissen

Wie zuvor angemerkt, ist neben der Tatsache, dass die Fälle in Knoten 9 jeweils etwa zur Hälfte in die beiden Kreditrating-Kategorien fallen, die Tatsache, dass die vorhergesagte Kategorie “gut” lautet, problematisch, wenn das Hauptziel darin besteht, ein Modell zu konstruieren, mit dem das Risiko für schlechtes Kreditrating korrekt identifiziert wird. Sie können zwar vielleicht nicht die Aussagekraft von Knoten 9 erhöhen, doch Sie können das Modell so verfeinern, dass die Quote für die richtige Klassifizierung der Fälle mit schlechtem Kreditrating erhöht wird. Beachten Sie jedoch, dass dies gleichzeitig zu einer höheren Fehlklassifizierungsquote für die Fälle mit gutem Kreditrating führt.

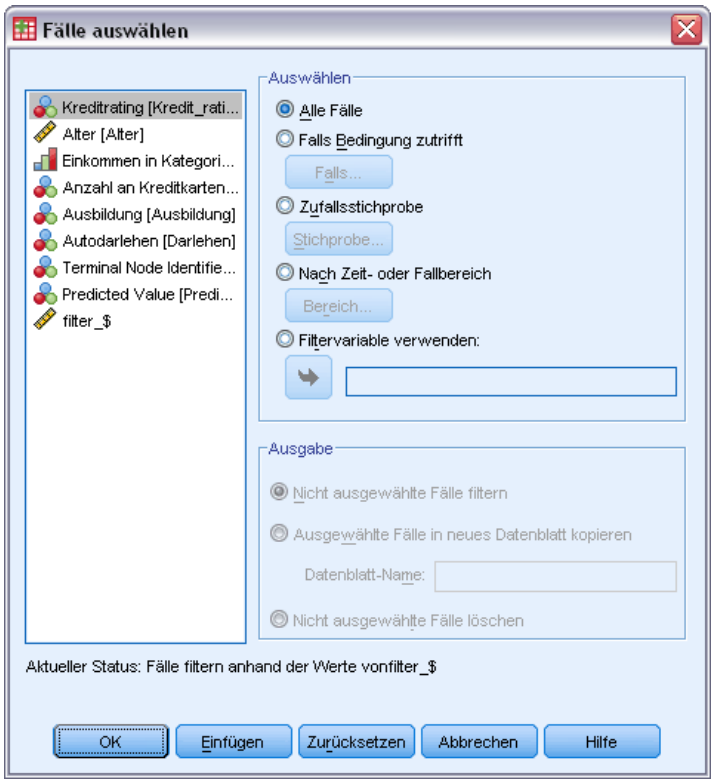
Zunächst müssen Sie die Fallfilterung deaktivieren, sodass wieder alle Fälle in der Analyse verwendet werden.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
  Fälle auswählen...

▶ Wählen Sie im Dialogfeld “Fälle auswählen” die Option Alle Fälle aus und klicken Sie anschließend auf OK.

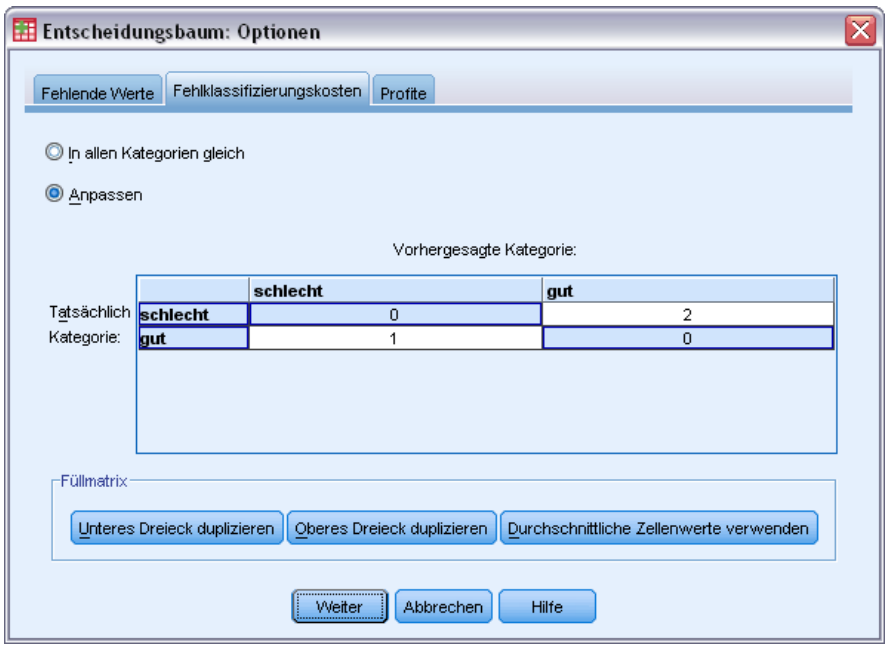
Abbildung 4-21
Dialogfeld “Fälle auswählen”



▶ Öffnen Sie noch einmal das Dialogfeld “Entscheidungsbaum” und klicken Sie auf Optionen.

▶ Klicken Sie auf die Registerkarte Fehlklassifizierungskosten.

Abbildung 4-22
Dialogfeld "Optionen", Registerkarte "Fehlklassifizierungskosten"

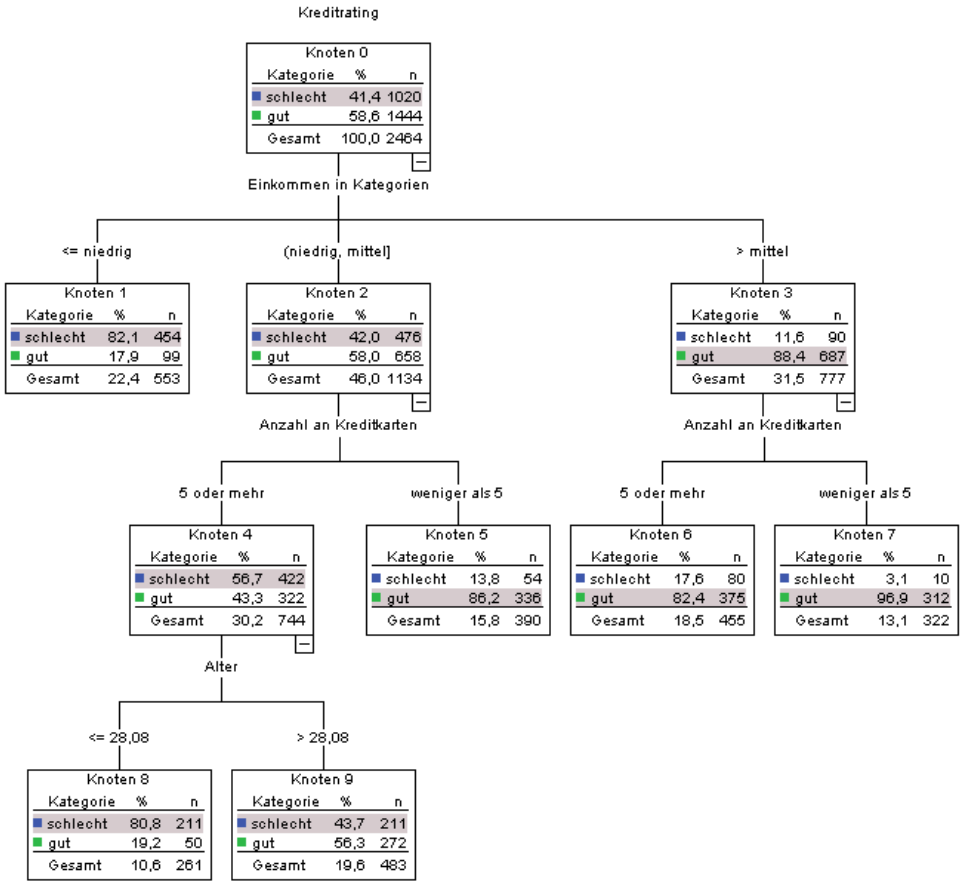


▶ Wählen Sie Benutzerdefiniert aus und geben Sie unter "Tatsächliche Kategorie *Schlecht*", "Vorhergesagte Kategorie *Gut*" den Wert 2 ein.

Dadurch werden die "Kosten" der falschen Klassifizierung eines schlechten Kreditrisikos als gut doppelt so hoch gewertet wie die "Kosten" der falschen Klassifizierung eines guten Kreditrisikos als schlecht.

▶ Klicken Sie auf Weiter und danach im Hauptdialogfeld auf OK, um die Prozedur auszuführen.

Abbildung 4-23
Baummodell mit angepassten Kostenwerten



Auf den ersten Blick sieht der durch die Prozedur erstellte Baum im Wesentlichen genauso aus wie der ursprüngliche Baum. Eine genauere Betrachtung ergibt jedoch, dass zwar die Verteilung der Fälle in den einzelnen Knoten gleich geblieben ist, sich jedoch einige vorhergesagte Kategorien geändert haben.

Bei den Endknoten bleiben die vorhergesagten Kategorien in allen Knoten gleich bis auf einen: Knoten 9. Die vorhergesagte Kategorie lautet nun *Schlecht*, obwohl sich etwas mehr als die Hälfte der Fälle in der Kategorie *Gut* befinden.

Da die Prozedur nun für die Fehlklassifizierung schlechter Kreditrisiken als gute Kreditrisiken höhere Kosten ansetzt, fällt nun jeder Knoten, in dem die Fälle ungefähr gleichmäßig auf die beiden Kategorien verteilt sind, in die vorhergesagte Kategorie *Schlecht*, selbst wenn sich eine leichte Mehrheit der Fälle in der Kategorie *Gut* befindet.

Diese Änderung in der vorhergesagten Kategorie ist auch in der Klassifikationstabelle zu sehen.

Abbildung 4-24
Risiko- und Klassifikationstabellen auf der Grundlage der angepassten Kosten

Risiko

Schätzer	Standardfehler
,288	,011

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Klassifikation

Beobachtet	Vorhergesagt		
	schlecht	gut	Prozent korrekt
schlecht	876	144	85,9%
gut	421	1023	70,8%
Gesamtprozentsatz	52,6%	47,4%	77,1%

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

- Fast 86 % der schlechten Kreditrisiken sind nun richtig klassifiziert, gegenüber vorher nur 65 %.
- Andererseits ist die korrekte Klassifizierung guter Kreditrisken von 90 % auf 71 % gesunken und der Gesamtwert für die korrekte Klassifizierung ist von 79,5 % auf 77,1 % gesunken.

Beachten Sie außerdem, dass der Risikoschätzer und die Gesamtquote für korrekte Klassifizierung nicht mehr zueinander konsistent sind. Bei einer Gesamtquote für korrekte Klassifizierung von 77,1 % wäre eigentlich ein Risikoschätzer von 0,229 zu erwarten. Durch die Erhöhung der Kosten für die Fehlklassifizierung von Fällen mit schlechtem Kreditrating wurde in diesem Beispiel der Risikowert erhöht, was seine Interpretation komplizierter macht.

Zusammenfassung

Mit Baummodellen können Sie Fälle in Gruppen einordnen, die durch bestimmte Merkmale identifiziert werden, beispielsweise die Merkmale, die Bankkunden mit guter oder schlechter Kredit-Historie zugeordnet werden können. Wenn ein bestimmtes vorhergesagtes Ergebnis wichtiger ist als andere mögliche Ergebnisse, können Sie das Modell verfeinern, um diesem Ergebnis höhere Fehlklassifizierungskosten zuzuordnen. Allerdings werden durch die Verringerung der Fehlklassifizierungsquoten für ein Ergebnis die Fehlklassifizierungsquoten für andere Ergebnisse erhöht.

Kapitel
5

Konstruieren eines Bewertungsmodells

Eine der leistungsstärksten und nützlichsten Funktionen der Prozedur “Entscheidungsbaum” ist die Fähigkeit, Modelle zu konstruieren, die dann auf andere Datendateien angewendet werden können, um Ergebnisse vorherzusagen. Beispielsweise können wir auf der Grundlage einer Datendatei, die sowohl demografische Informationen als auch Informationen zu Fahrzeugverkaufspreisen enthält, ein Modell erstellen, mit dem vorhergesagt werden kann, welchen Betrag Personen mit ähnlichen demografischen Merkmalen wahrscheinlich für ein neues Auto ausgeben, und das Modell anschließend auf andere Datendateien anwenden, in denen demografische Daten vorhanden sind, jedoch keine Informationen über frühere Fahrzeugkäufe.

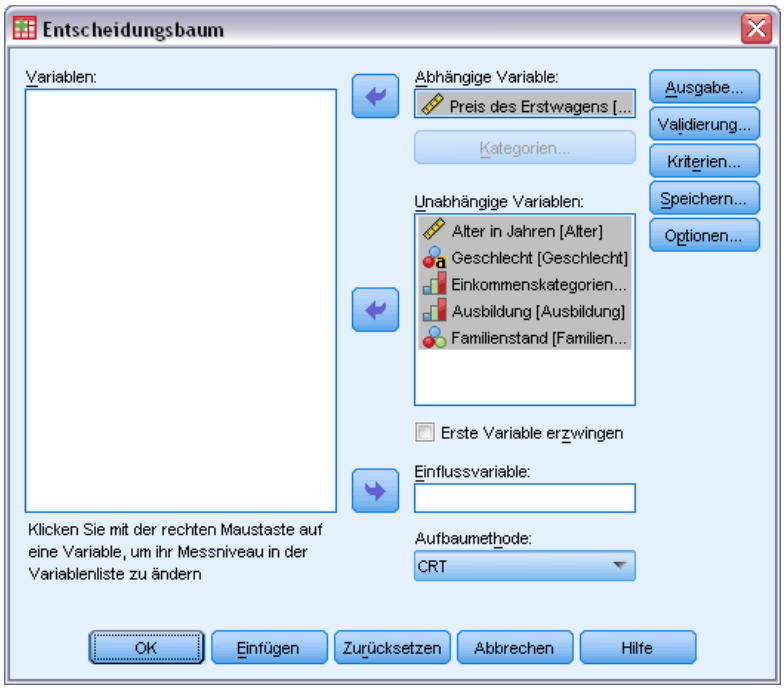
In diesem Beispiel wird die Datendatei *tree_car.sav* verwendet. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Decision Trees (Entscheidungsbäume) 18*.

Konstruieren des Modells

- Zum Erstellen einer Entscheidungsbaum-Analyse wählen Sie die folgenden Befehle aus den Menüs aus:

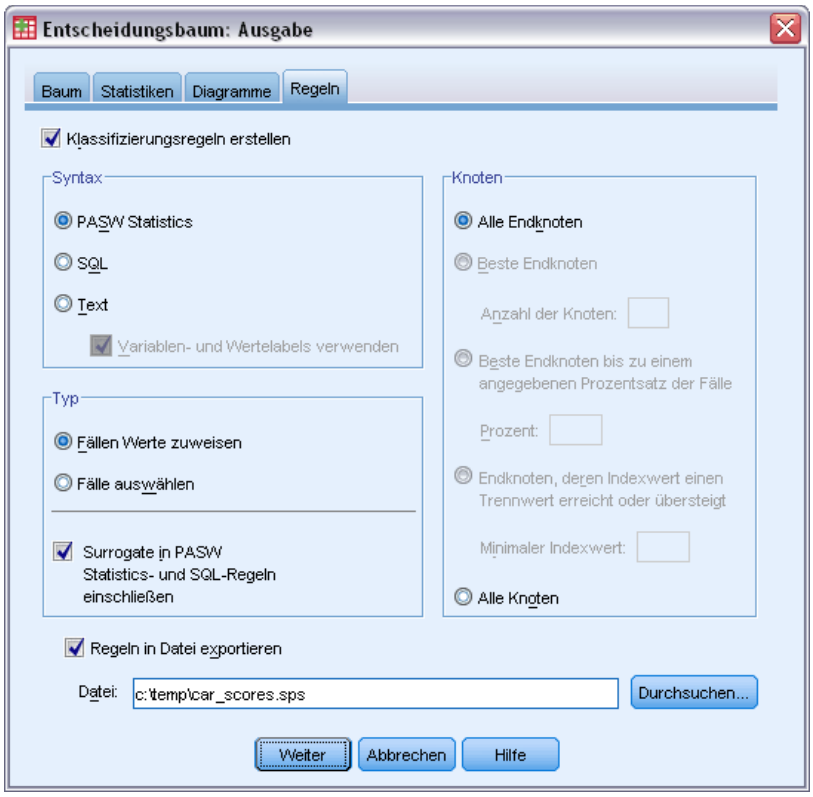
Analysieren
  Klassifizieren
    Baum...

Abbildung 5-1
Dialogfeld "Entscheidungsbaum"



- Wählen Sie *Preis des Erstwagens* als abhängige Variable aus.
- Wählen Sie alle verbleibenden Variablen als unabhängige Variablen aus. (Die Prozedur schließt automatisch alle Variablen aus, die keinen signifikanten Beitrag zum endgültigen Modell leisten.)
- Wählen Sie als Aufbaumethode CRT aus.
- Klicken Sie auf Ausgabe.

Abbildung 5-2
Dialogfeld "Ausgabe", Registerkarte "Regeln"



- Klicken Sie auf die Registerkarte Regeln.
- Aktivieren Sie Klassifizierungsregeln erstellen.
- Wählen Sie für "Syntax" PASW Statistics.
- Wählen Sie als Typ Fällen Wert zuweisen aus.
- Aktivieren Sie Regeln in Datei exportieren und geben Sie einen Dateinamen und eine Verzeichnisposition ein.

  Merken Sie sich den Dateinamen und die Verzeichnisposition oder schreiben Sie sie auf, da Sie diese Angaben bald wieder benötigen. Wenn Sie keinen Verzeichnispfad angeben, wissen Sie möglicherweise nicht, wo die Datei gespeichert wurde. Mit der Schaltfläche Durchsuchen können Sie zu einer bestimmten (gültigen) Verzeichnisposition wechseln.
- Klicken Sie auf Weiter und anschließend auf OK, um die Prozedur auszuführen und das Baummodell zu konstruieren.

Bewertung des Modells

Bevor Sie das Modell auf andere Datendateien anwenden, sollten Sie sicherstellen, dass das Modell gut mit den ursprünglichen Daten, die für die Modellkonstruktion verwendet wurden, arbeitet.

Modellübersicht

Abbildung 5-3
Modellzusammenfassungstabelle

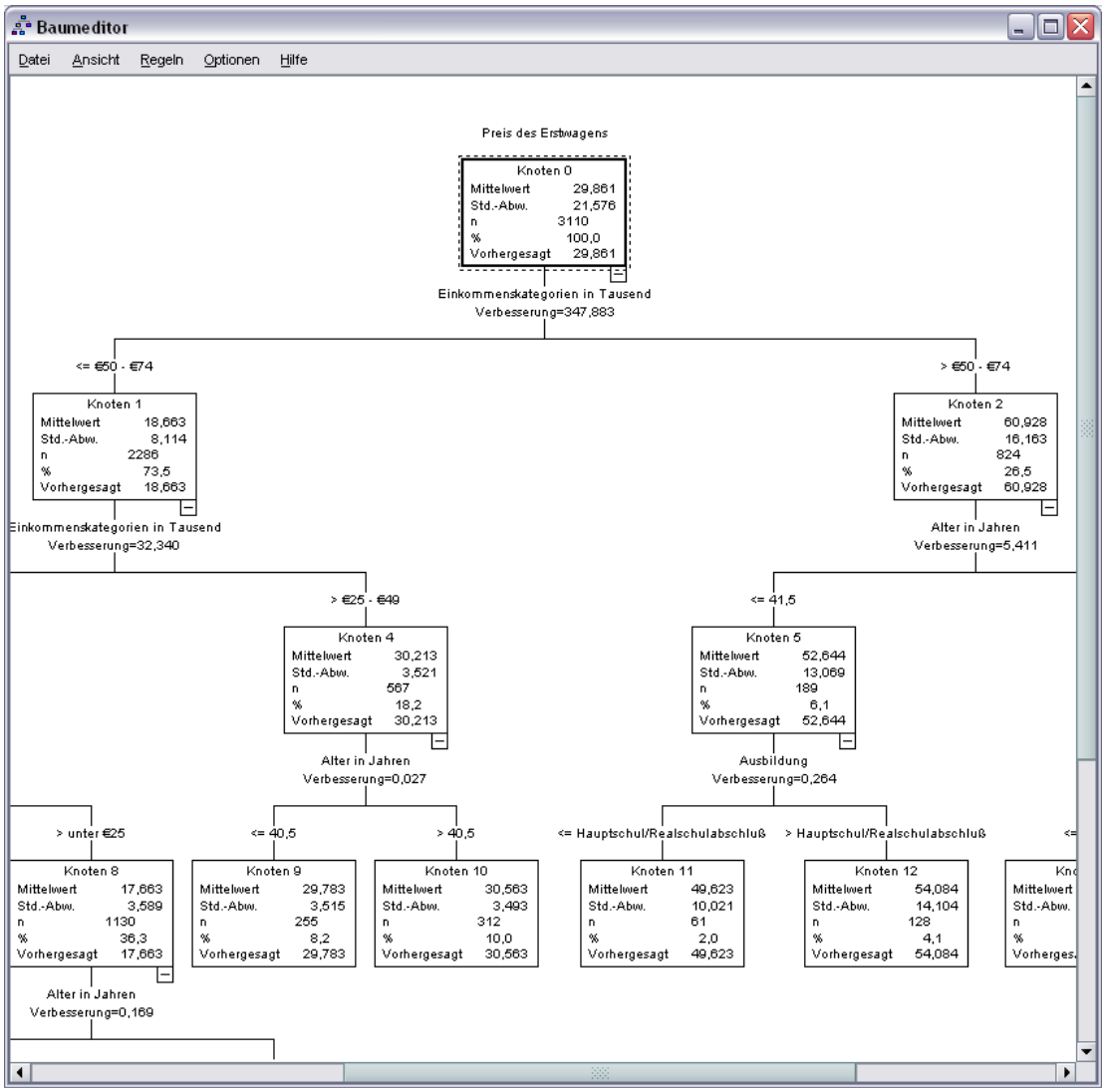
Spezifikationen	Aufbaumethode	CRT
	Abhängige Variable	Preis des Erstwagens
	Unabhängige Variablen	Alter in Jahren, Geschlecht, Einkommenskategorien in Tausend, Ausbildung, Familienstand
	Validierung	NONE
	Maximale Baumtiefe	5
	Mindestanzahl der Fälle im übergeordneten Knoten	100
	Mindestanzahl der Fälle im untergeordneten Knoten	50
Ergebnisse	Aufgenommene unabhängige Variablen	Einkommenskategorien in Tausend, Alter in Jahren, Ausbildung
	Anzahl der Knoten	29
	Anzahl der Endknoten	15
	Tiefe	5

Die Modellzusammenfassungstabelle zeigt an, dass nur drei der ausgewählten unabhängigen Variablen einen Beitrag leisteten, der signifikant genug ist, dass ihre Aufnahme in das endgültige Modell gerechtfertigt ist: *einkomme*, *alter* und *ausbildu*. Diese Informationen sind wichtig, wenn Sie das Modell auf andere Datendateien anwenden möchten, da die im Modell verwendeten unabhängigen Variablen in allen Datendateien vorhanden sein müssen, auf die das Modell angewendet werden soll.

Die Zusammenfassungstabelle zeigt außerdem an, dass das Baummodell selbst offenbar nicht besonders einfach ist, da es 29 Knoten und 15 Endknoten aufweist. Das ist möglicherweise kein Problem, wenn Sie ein zuverlässiges Modell wünschen, das der praktischen Anwendung dienen soll, und nicht ein einfaches Modell, das einfach zu beschreiben oder zu erklären ist. Natürlich sollte sich das Modell aus Gründen der Praktikabilität nicht auf zu viele unabhängige (Einfluss-)Variablen stützen. In diesem Fall ist das kein Problem, da nur drei unabhängige Variablen im endgültigen Modell enthalten sind.

Baummodelldiagramm

Abbildung 5-4
Baummodelldiagramm im Baum-Editor

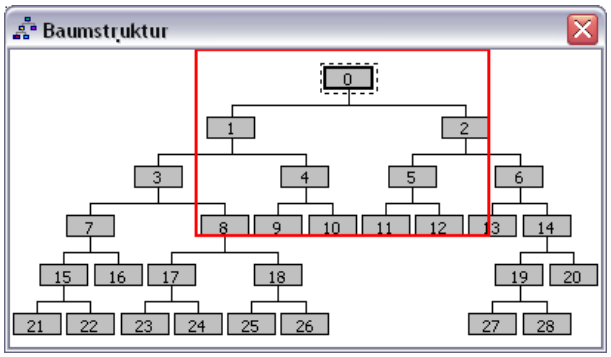


Das Baummodelldiagramm enthält so viele Knoten, dass es schwierig ist, das ganze Modell auf einmal in einer Größe anzuzeigen, in der die Informationen zum Knoteninhalt noch lesbar sind. Mithilfe der Baumstruktur können Sie den gesamten Baum anzeigen:

- Doppelklicken Sie auf den Baum im Viewer, um den Baum-Editor zu öffnen.
- Wählen Sie die folgenden Menübefehle des Baum-Editors aus:

Ansicht
Baumstruktur

Abbildung 5-5
Baumstruktur



- In der Baumstruktur wird der gesamte Baum angezeigt. Sie können die Größe des Fensters "Baumstruktur" ändern, wodurch die Strukturanzeige des Baums entsprechend der Fenstergröße vergrößert bzw. verkleinert wird.
- Der markiert Bereich in der Baumstruktur ist der Bereich des Baums, der derzeit im Baum-Editor angezeigt wird.
- Mithilfe der Baumstruktur können Sie im Baum navigieren und Knoten auswählen:

Für weitere Informationen siehe Baumstruktur in Kapitel 2 auf S. 43.

Bei abhängigen metrischen Variablen zeigt jeder Knoten den Mittelwert und die Standardabweichung der abhängigen Variablen an. Knoten 0 zeigt einen Gesamtmittelwert für den Fahrzeugkaufpreis von ca. 29,9 (in Tausend) an, mit einer Standardabweichung von ca. 21,6.

- Knoten 1, der für Fälle mit einem Einkommen von weniger als 75 (ebenfalls in Tausend) steht, weist einen mittleren Fahrzeugpreis von nur 18,7 auf.
- Knoten 2 dagegen, der für Fälle mit einem Einkommen von mindestens 75 steht, weist einen mittleren Fahrzeugpreis von 60,9 auf.

Eine eingehendere Untersuchung des Baums würde zeigen, dass *alter* und *ausbildu* ebenfalls eine Beziehung zum Fahrzeugkaufpreis aufweisen; im Moment interessieren wir uns jedoch in erster Linie für die praktische Anwendung des Modells und weniger für eine detaillierte Untersuchung seiner Komponenten.

Risikoschätzer

Abbildung 5-6
Risikotabelle

Risiko

Schätzer	Standardfehler
68,485	2,985

Aufbaumethode: CRT
Abhängige Variable: Preis des Erstwagens

Keines der Ergebnisse, die wir bisher untersucht haben, deutet darauf hin, dass dies ein besonders gutes Modell ist. Ein Indikator für die Leistungsfähigkeit eines Modells ist der Risikoschätzer. Bei einer abhängigen metrischen Variablen ist der Risikoschätzer ein Maß für die Varianz innerhalb des Knotens, was für sich genommen noch nicht sehr aussagekräftig ist. Eine niedrigere Varianz weist auf ein besseres Modell hin, doch die Varianz ist relativ zur Maßeinheit. Wenn der Preis beispielsweise nicht in Tausend angegeben worden wäre, wäre der Risikoschätzer um ein Tausendfaches größer.

Um bei einer abhängigen metrischen Variablen eine sinnvolle Interpretation für den Risikoschätzer zu erarbeiten, muss ein gewisser Aufwand betrieben werden:

- Die Gesamtvarianz ist gleich der (Fehler-)Varianz innerhalb der einzelnen Knoten plus der (erklärten) Varianz zwischen den Knoten.
- Die Varianz innerhalb der Knoten ist der Wert für den Risikoschätzer: 68.485.
- Die Gesamtvarianz ist die Varianz für die abhängigen Variablen vor der Berücksichtigung von unabhängigen Variablen, nämlich die Varianz am Stammknoten.
- Die am Stammknoten angezeigte Standardabweichung beträgt 21,576; also ist die Gesamtvarianz das Quadrat dieses Werts: 465.524.
- Der Anteil der Varianz der auf Fehler zurückzuführen ist (unerklärte Varianz) beträgt 68,485/465,524 = 0,147.
- Der Anteil der von diesem Modell erklärten Varianz beträgt 1–0,147 = 0,853 bzw. 85,3 %, was anzeigt, dass es sich um ein ziemlich gutes Modell handelt. (Eine ähnliche Interpretation wie die Gesamtquote für die korrekte Klassifizierung für eine abhängige kategoriale Variable.)

Anwenden des Modells auf eine andere Datendatei

Nachdem wir festgestellt haben, dass das Modell eine angemessene Qualität aufweist, können wir das Modell nun auf andere Datendateien mit ähnlichen Variablen vom Typ *alter*, *einkomme* und *ausbildu* anwenden und eine neue Variable erstellen, die für jeden Fall in dieser Datei den vorhergesagten Kaufpreis angibt. Dieser Prozess wird häufig als **Bewertung** bezeichnet.

Bei der Erstellung des Modells haben wir angegeben, dass "Regeln" für die Zuweisung von Werten zu Fällen in einer Textdatei (in Form von Befehlssyntax) gespeichert werden sollen. Wir verwenden nun die Befehle in dieser Datei, um Werte in einer anderen Datei zu erstellen.

▶ Öffnen Sie die Daten-Datei *tree_score_car.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Decision Trees (Entscheidungsbäume) 18*.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

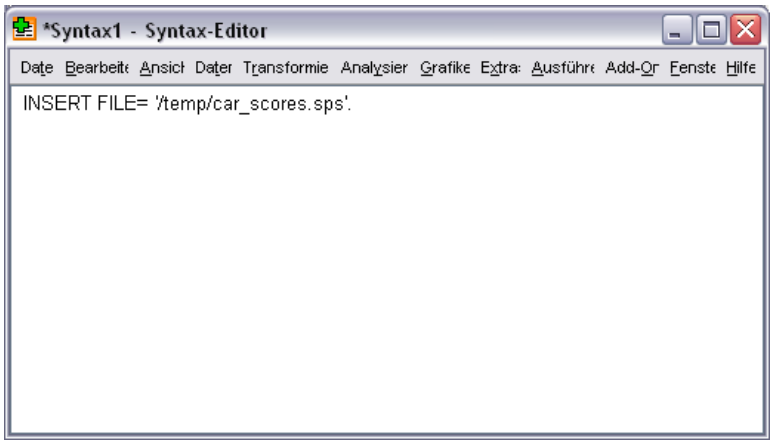
Datei
  Neu
    Syntax

▶ Geben Sie im Fenster für die Befehlssyntax Folgendes ein:

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

Wenn Sie einen anderen Dateinamen oder einen anderen Speicherort verwendet haben, müssen Sie die Eingabe entsprechend abwandeln.

Abbildung 5-7
Syntax-Fenster mit Befehl INSERT zur Ausführung einer Befehlsdatei

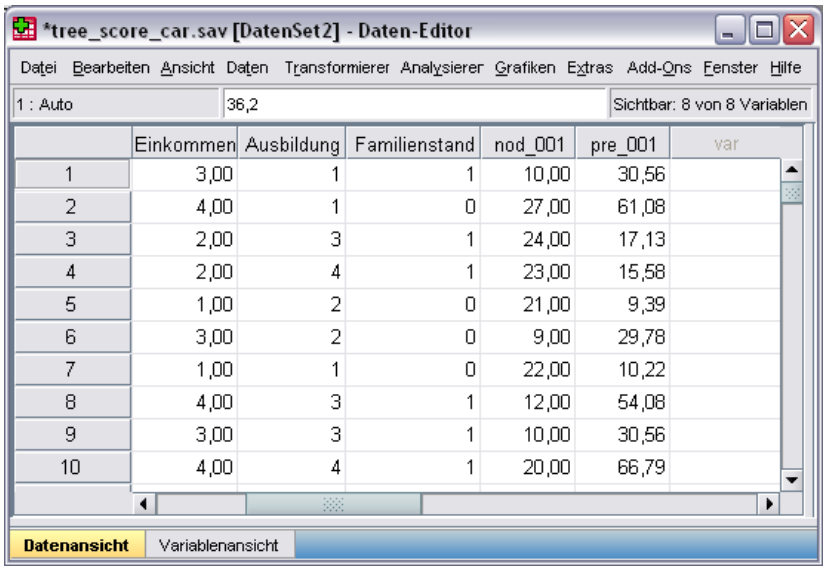

*Syntax1 - Syntax-Editor
Date Bearbeite Ansicl Dater Transformie Analysier Grafike Extra: Ausführe Add-On Fenste Hilfe
INSERT FILE= '/temp/car_scores.sps'.

Der Befehl `INSERT` führt die Befehle in der angegebenen Datei aus, nämlich der “Regel”-Datei, die bei der Erstellung des Modells angelegt wurde.

▶ Wählen Sie die folgenden Befehle aus den Menüs des Befehlssyntax-Fensters aus:

Ausführen
Alle

Abbildung 5-8
Zur Datendatei hinzugefügte vorhergesagte Werte


*tree_score_car.sav [DatenSet2] - Daten-Editor
Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Add-Ons Fenster Hilfe
1 : Auto
36,2
Sichtbar: 8 von 8 Variablen
Einkommen Ausbildung Familienstand nod_001 pre_001 var
1 3,00 1 1 10,00 30,56
2 4,00 1 0 27,00 61,08
3 2,00 3 1 24,00 17,13
4 2,00 4 1 23,00 15,58
5 1,00 2 0 21,00 9,39
6 3,00 2 0 9,00 29,78
7 1,00 1 0 22,00 10,22
8 4,00 3 1 12,00 54,08
9 3,00 3 1 10,00 30,56
10 4,00 4 1 20,00 66,79
Datenansicht
Variablenansicht

Dadurch werden zwei neue Variablen zu der Datendatei hinzugefügt:

- *nod_001* enthält die vom Modell für die einzelnen Fälle vorhergesagten Endknotennummern.
- *pre_001* enthält den vorhergesagten Wert für den Fahrzeugkaufpreis für die einzelnen Fälle.

Da Regeln für die Zuweisung von Werten für Endknoten angefordert wurden, stimmt die Anzahl der möglichen vorausgesagten Werte mit der Anzahl der Endknoten überein, in diesem Fall 15. So weist beispielsweise jeder Fall mit einer vorhergesagten Knotennummer von 10 denselben vorhergesagten Fahrzeugkaufpreis auf: 30.56. Dies ist – und zwar nicht zufällig – der für den Endknoten 10 im ursprünglichen Modell angegebene Mittelwert.

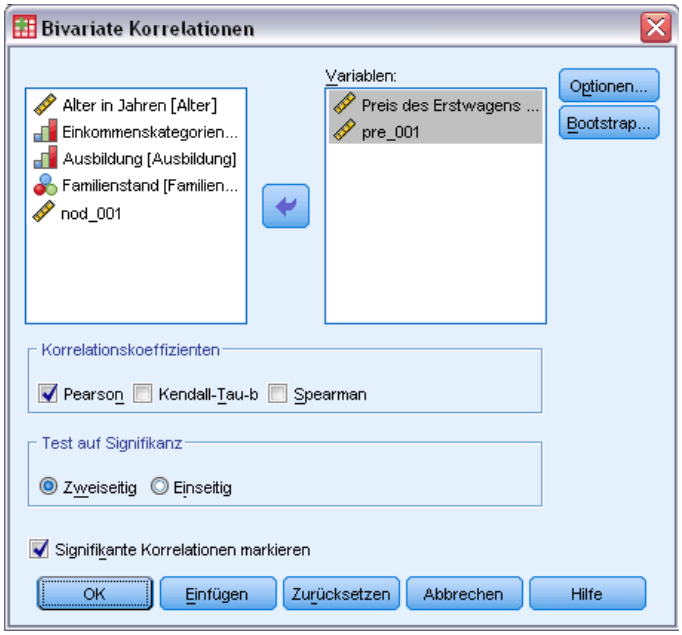
Normalerweise würden Sie zwar das Modell auf Daten anwenden, bei denen der Wert der abhängigen Variablen nicht bekannt ist, in diesem Beispiel jedoch enthält die Datendatei, auf die das Modell angewendet wird, diese Informationen, sodass Sie die Modellvorhersagen mit den tatsächlichen Werten vergleichen können.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
Korrelation
Bivariat...

▶ Wählen Sie *Preis des Erstwagens* und *pre_001* als abhängige Variablen aus.

Abbildung 5-9
Dialogfeld "Bivariate Korrelationen"



▶ Klicken Sie auf OK, um die Prozedur auszuführen.

Abbildung 5-10
Korrelation zwischen tatsächlichem und vorhergesagtem Fahrzeugpreis

		Preis des Erstwagens	pre_001
Preis des Erstwagens	Korrelation nach Pearson	1	,919**
	Signifikanz (2-seitig)		,000
	N	3290	3290
pre_001	Korrelation nach Pearson	,919**	1
	Signifikanz (2-seitig)	,000	
	N	3290	3290

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Die Korrelation von 0,92 weist auf eine sehr hohe positive Korrelation zwischen tatsächlichem und vorhergesagtem Fahrzeugpreis auf, die anzeigt, dass das Modell gut funktioniert.

Zusammenfassung

Mit der Prozedur “Entscheidungsbaum” können Sie Modelle konstruieren, die dann auf andere Datendateien angewendet werden können, um Ergebnisse vorherzusagen. Die Zieldatendatei muss Variablen mit demselben Namen enthalten wie die im endgültigen Modell enthaltenen unabhängigen Variablen, die mit derselben Metrik gemessen werden und die dieselben benutzerdefiniert fehlenden Werte aufweisen (sofern vorhanden). In der Zieldatendatei müssen jedoch weder die abhängige Variable noch die aus dem endgültigen Modell ausgeschlossenen unabhängigen Variablen enthalten sein.

Kapitel 6

Fehlende Werte in Baummodellen

Bei den unterschiedlichen Aufbaumethoden werden fehlende Werte für unabhängige Variablen (Einflußvariablen) auf verschiedene Weise behandelt:

- Bei CHAID und Exhaustive CHAID werden alle system- und benutzerdefiniert fehlenden Werte für die einzelnen unabhängigen Variablen als einzige Kategorie behandelt. Bei metrischen und ordinalen unabhängigen Variablen wird diese Kategorie ggf. anschließend mit anderen Kategorien dieser unabhängigen Variable zusammengeführt, je nach den Aufbaukriterien.
- Bei CRT und QUEST werden nach Möglichkeit **Surrogate** für unabhängige Variablen (Einflußvariablen) verwendet. In Situationen, in denen der Wert für die betreffende Variable fehlt, werden andere unabhängige Variablen, die einen hohen Grad an Zusammenhang mit der ursprünglichen Variable besitzen, zur Klassifizierung herangezogen. Diese alternativen Einflussvariablen werden als Surrogate bezeichnet.

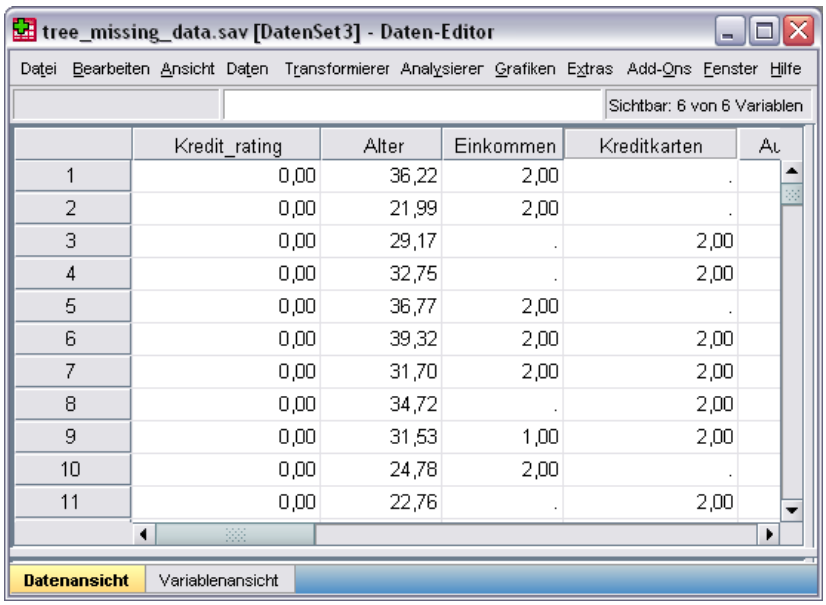
Dieses Beispiel verdeutlicht die Unterschiede zwischen CHAID und CRT, wenn Werte für unabhängige Variablen im Modell fehlen.

In diesem Beispiel wird die Datendatei *tree_missing_data.sav* verwendet. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Decision Trees (Entscheidungsbäume) 18*.

Hinweis: Bei nominalen unabhängigen Variablen und nominalen abhängigen Variablen können Sie angeben, dass **benutzerdefiniert fehlende** Werte als gültige Werte behandelt werden sollen. Die Werte werden somit wie andere, nichtfehlende Werte behandelt. Für weitere Informationen siehe Fehlende Werte definieren in Kapitel 1 auf S. 23.

Fehlende Werte bei CHAID

Abbildung 6-1
Kreditdaten mit fehlenden Werten


tree_missing_data.sav [DatenSet3] - Daten-Editor
Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Add-Ons Fenster Hilfe
Sichtbar: 6 von 6 Variablen
Datenansicht
Variablenansicht

	Kredit_rating	Alter	Einkommen	Kreditkarten	Au
1	0,00	36,22	2,00	.	
2	0,00	21,99	2,00	.	
3	0,00	29,17	.	2,00	
4	0,00	32,75	.	2,00	
5	0,00	36,77	2,00	.	
6	0,00	39,32	2,00	2,00	
7	0,00	31,70	2,00	2,00	
8	0,00	34,72	.	2,00	
9	0,00	31,53	1,00	2,00	
10	0,00	24,78	2,00	.	
11	0,00	22,76	.	2,00	

Wie beim Beispiel für das Kreditrisiko (weitere Informationen finden Sie unter Kapitel 4) wird auch in diesem Beispiel ein Modell erstellt, mit dem hohe und niedrige Kreditrisiken ermittelt werden sollen. Der wichtigste Unterschied liegt darin, dass diese Datendatei fehlende Werte für einige unabhängige Variablen im Modell aufweist.

- Zum Erstellen einer Entscheidungsbaum-Analyse wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Klassifizieren
    Baum...

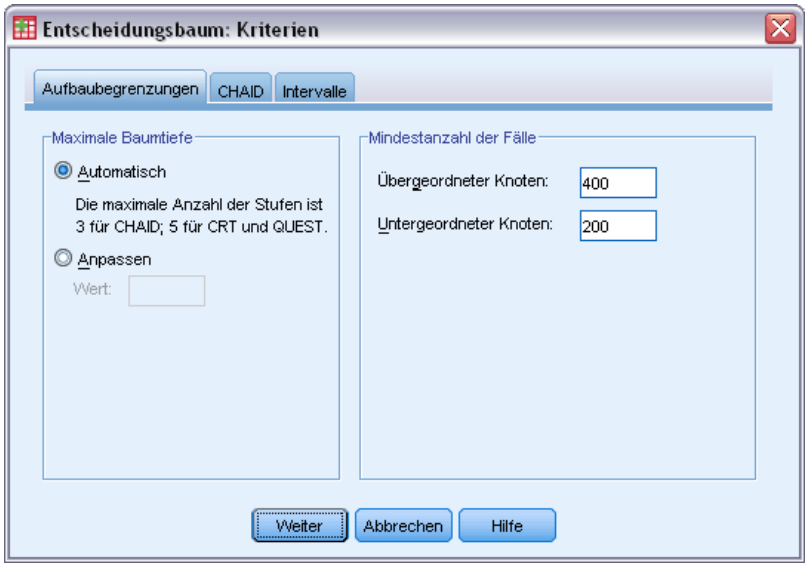
Abbildung 6-2
Dialogfeld "Entscheidungsbaum"

- Wählen Sie *Kreditrating* als abhängige Variable aus.
- Wählen Sie alle verbleibenden Variablen als unabhängige Variablen aus. (Die Prozedur schließt automatisch alle Variablen aus, die keinen signifikanten Beitrag zum endgültigen Modell leisten.)
- Wählen Sie als Aufbaumethode die Option CHAID.

  Der Baum soll in diesem Beispiel relativ einfach gehalten werden. Der Aufbau des Baums wird daher eingeschränkt, indem eine höhere Mindestanzahl der Fälle für die über- und untergeordneten Knoten angegeben wird.

- Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf Kriterien.

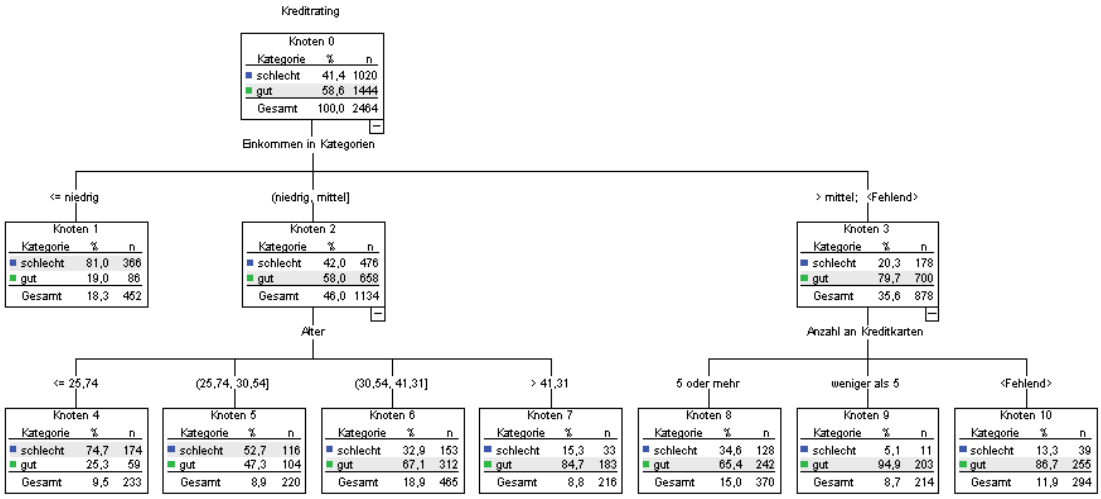
Abbildung 6-3
Dialogfeld "Kriterien", Registerkarte "Aufbaubegrenzungen"



- Geben Sie unter "Mindestanzahl der Fälle" den Wert 400 für "Übergeordneter Knoten" sowie den Wert 200 für "Untergeordneter Knoten" ein.
- Klicken Sie auf Weiter und dann auf OK. Die Prozedur wird gestartet.

CHAID-Ergebnisse

Abbildung 6-4
CHAID-Baum mit fehlenden Werten für unabhängige Variablen



Bei Knoten 3 wird der Wert für *Einkommen in Kategorien* als *>Mittel;<fehlend>* aufgeführt. Der Knoten enthält also Fälle in der Kategorie mit hohem Einkommen und außerdem Fälle mit fehlenden Werten für *Einkommen in Kategorien*.

Der Endknoten 10 enthält Fälle mit fehlenden Werten für *Anzahl an Kreditkarten*. Bei der Ermittlung risikoloser Kredite ist dieser Endknoten am zweitbesten geeignet; wenn dieses Modell zur Vorhersage risikoloser Kredite dienen soll, kann dies zu Problemen führen. Ein Modell, das einen risikolosen Kredit vorhersagt, ist nutzlos, wenn nicht bekannt ist, wie viele Kreditkarten der Kunde besitzt und womöglich auch die Angaben zur Einkommenshöhe in einigen Fällen fehlen.

Abbildung 6-5
Risiko- und Klassifizierungstabellen für das CHAID-Modell

Risiko

Schätzer	Standardfehler
,249	,009

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Klassifikation

	Vorhergesagt		
Beobachtet	schlecht	gut	Prozent korrekt
schlecht	656	364	64,3%
gut	249	1195	82,8%
Gesamtprozentsatz	36,7%	63,3%	75,1%

Aufbaumethode: CHAID
Abhängige Variable: Kreditrating

Die Risiko- und Klassifizierungstabellen weisen darauf hin, dass das CHAID-Modell etwa 75 % der Fälle korrekt klassifiziert. Dieses Ergebnis ist zwar nicht schlecht, aber noch lange nicht gut. Außerdem besteht Grund zur Annahme, dass die Rate der richtigen Klassifizierung für risikolose Kreditfälle zu optimistisch sein könnte, weil diese Rate teilweise auf der willkürlichen Annahme beruht, dass fehlende Daten für zwei unabhängige Variablen (*Einkommen in Kategorien* und *Anzahl an Kreditkarten*) ein Anzeichen für einen risikolosen Kredit sind.

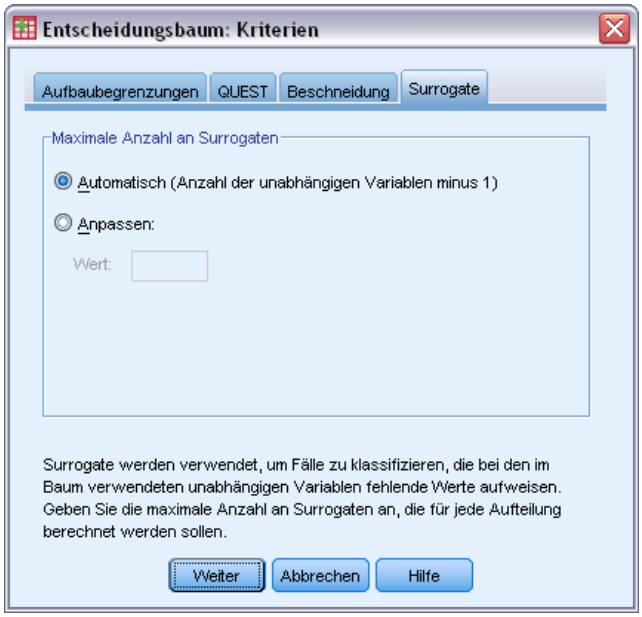
Fehlende Werte bei CRT

Wiederholen Sie nun dieselbe grundlegende Analyse und verwenden Sie dabei die Aufbaumethode CRT.

- Wählen Sie im Hauptdialogfeld “Entscheidungsbaum” als Aufbaumethode die Option CRT.
- Klicken Sie auf Kriterien.
- Stellen Sie sicher, dass die Mindestanzahl der Fälle weiterhin 400 für übergeordnete Knoten bzw. 200 für untergeordnete Knoten beträgt.
- Klicken Sie auf die Registerkarte Surrogate.

Hinweis: Die Registerkarte “Surrogate” ist nur dann sichtbar, wenn Sie die Aufbaumethode CRT oder QUEST verwenden.

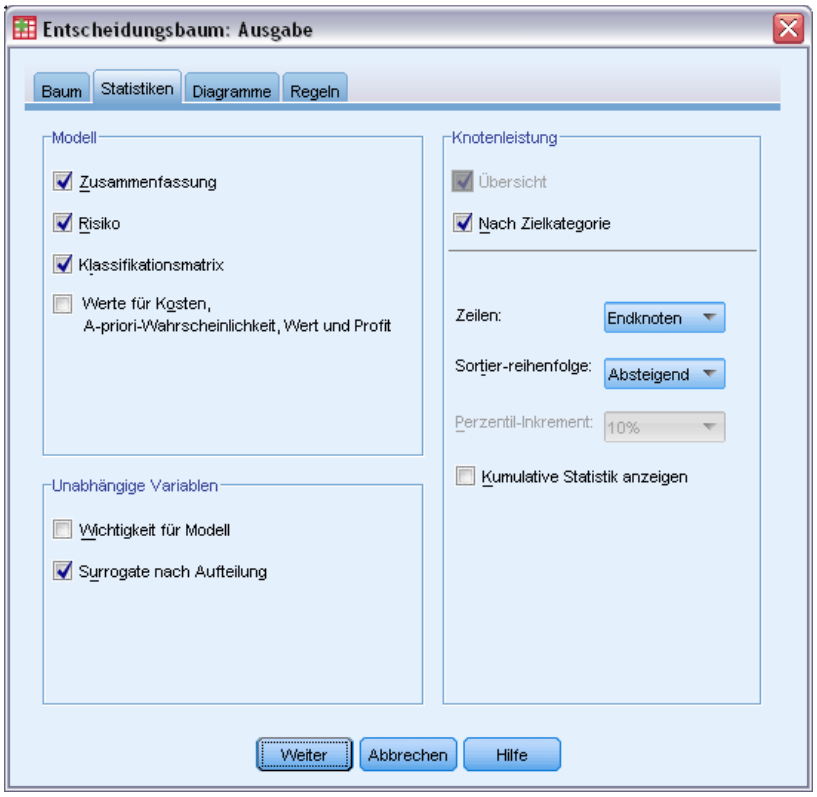
Abbildung 6-6
Dialogfeld "Kriterien", Registerkarte "Surrogate"


Entscheidungsbaum: Kriterien
Aufbaubegrenzungen
QUEST
Beschneidung
Surrogate
Maximale Anzahl an Surrogaten
Automatisch (Anzahl der unabhängigen Variablen minus 1)
Anpassen:
Wert:
Surrogate werden verwendet, um Fälle zu klassifizieren, die bei den im Baum verwendeten unabhängigen Variablen fehlende Werte aufweisen. Geben Sie die maximale Anzahl an Surrogaten an, die für jede Aufteilung berechnet werden sollen.
Weiter
Abbrechen
Hilfe

Mit der Einstellung Automatisch wird bei jeder Knotenaufteilung für die unabhängige Variable geprüft, ob eine andere unabhängige Variable im Modell als Surrogat infrage kommt. Dieses Beispiel enthält nur wenige unabhängige Variablen; die Einstellung Automatisch ist daher ohne weiteres möglich.

- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Entscheidungsbaum" auf Ausgabe.

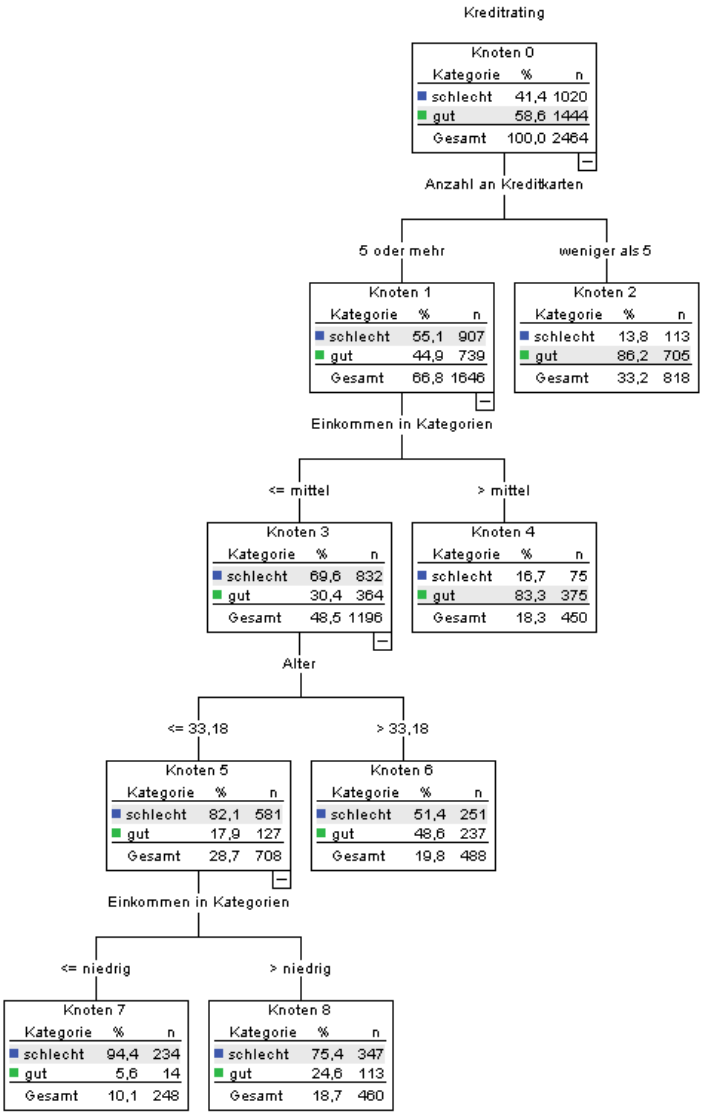
Abbildung 6-7
Dialogfeld "Optionen", Registerkarte "Statistik"



- Klicken Sie auf die Registerkarte Statistik.
- Wählen Sie Surrogate nach Aufteilung.
- Klicken Sie auf Weiter und dann auf OK. Die Prozedur wird gestartet.

CRT-Ergebnisse

Abbildung 6-8
CRT-Baum mit fehlenden Werten für unabhängige Variablen



Auf den ersten Blick ist ersichtlich, dass dieser Baum keine große Ähnlichkeit mit dem CHAID-Baum besitzt. Für sich allein betrachtet hat dies noch nicht viel zu bedeuten. In einem CRT-Baummodell sind alle Aufteilungen binär; jeder übergeordnete Knoten wird also in nur zwei untergeordnete Knoten aufgeteilt. In einem CHAID-Modell können die übergeordneten Knoten in zahlreiche untergeordnete Knoten aufgeteilt werden. Die Bäume sehen also häufig anders aus, auch wenn sie dasselbe zugrundeliegende Modell darstellen.

Es gibt allerdings eine Reihe wichtiger Unterschiede:

- Die wichtigste unabhängige Variable (Einflussvariable) im CRT-Modell ist *Anzahl an Kreditkarten*, im CHAID-Modell dagegen *Einkommen in Kategorien*.

- Bei Fällen mit weniger als fünf Kreditkarten ist *Anzahl an Kreditkarten* die einzige relevante Einflussvariable für das Kreditrating und Knoten 2 ist ein Endknoten.
- Wie beim CHAID-Modell sind auch die Variablen *Einkommen in Kategorien* und *Alter* in diesem Modell vorhanden; *Einkommen in Kategorien* fungiert jedoch nun nicht mehr als erste Einflussvariable, sondern als zweite.
- Es liegen keine Knoten mit der Kategorie <*fehlend*> vor, weil bei CRT keine fehlenden Werte im Modell zugelassen, sondern Surrogat-Einflussvariablen verwendet werden.

Abbildung 6-9
Risiko- und Klassifizierungstabellen für das CRT-Modell

Risiko

Schätzer	Standardfehler
,224	,008

Aufbaumethode: CRT
Abhängige Variable: Kreditrating

Klassifikation

	Vorhergesagt		
Beobachtet	schlecht	gut	Prozent korrekt
schlecht	832	188	81,6%
gut	364	1080	74,8%
Gesamtprozentsatz	48,5%	51,5%	77,6%

Aufbaumethode: CRT
Abhängige Variable: Kreditrating

- Die Risiko- und Klassifizierungstabellen zeigen eine Gesamtrate für die korrekte Klassifizierung von nahezu 78 %, also eine leichte Verbesserung gegenüber dem CHAID-Modell (75 %).
- Die Rate für die richtige Klassifizierung risikobehafteter Kredite ist beim CRT-Modell deutlich höher: 81,6 % im Vergleich zu nur 64,3 % im CHAID-Modell.
- Die Rate für die korrekte Klassifizierung der Fälle mit gutem Kreditrating ist allerdings von 82,8 % beim CHAID-Modell auf 74,8 % beim CRT-Modell gefallen.

Surrogate

Die Unterschiede zwischen CHAID- und CRT-Modell liegen teilweise an der Verwendung von Surrogaten im CRT-Modell. Die Tabelle der Surrogate zeigt, wie die Surrogate im Modell genutzt wurden.

Abbildung 6-10
Tabelle der Surrogate

Übergeordneter Knoten	Unabhängige Variable		Verbesserung	Assoziation
0	Primär	Anzahl an Kreditkarten	,090	
	Surrogate	Autodarlehen	,052	,643
		Alter	,001	,004
1	Primär	Einkommen in Kategorien	,071	
	Surrogate	Alter	,001	,004
3	Primär	Alter	,022	
5	Primär	Einkommen in Kategorien	,006	
	Surrogate	Alter	3,93E-005	,009

Growing Method: CRT
Dependent Variable: Kreditrating

- Am Stammknoten (Knoten 0) ist *Anzahl an Kreditkarten* die beste unabhängige Variable (Einflussvariable).
- Bei allen Fällen mit fehlenden Werten für *Anzahl an Kreditkarten* wird *Autodarlehen* als Surrogat-Einflussvariable herangezogen, weil diese Variable relativ stark (0,643) mit *Anzahl an Kreditkarten* verbunden ist.
- Weist ein Fall auch einen fehlenden Wert für *Autodarlehen* auf, wird *Alter* als Surrogat verwendet (auch wenn hier nur ein äußerst geringer Wert von 0,004 für den Zusammenhang besteht).
- *Alter* wird außerdem als Surrogat für *Einkommen in Kategorien* in den Knoten 1 und 5 verwendet.

Zusammenfassung

Bei den verschiedenen Aufbaumethoden werden fehlende Daten auf unterschiedliche Weise behandelt. Wenn die Daten, aus denen das Modell erstellt wurde, zahlreiche fehlende Werte aufweisen (oder wenn Sie das Modell auf andere Datendateien anwenden möchten, bei denen viele Werte fehlen), sollten Sie die Auswirkungen der fehlenden Werte auf die verschiedenen Modelle überprüfen. Sollen fehlende Werte im Modell durch Surrogate ausgeglichen werden, verwenden Sie die Methode CRT oder QUEST.

Anhang

A

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses. Für jeder der folgenden Sprachen gibt es einen eigenen Ordner innerhalb des Unterverzeichnisses “Samples”: Englisch, Französisch, Deutsch, Italienisch, Japanisch, Koreanisch, Polnisch, Russisch, Vereinfachtes Chinesisch, Spanisch und Traditionelles Chinesisch.

Nicht alle Beispieldateien stehen in allen Sprachen zur Verfügung. Wenn eine Beispieldatei nicht in einer Sprache zur Verfügung steht, enthält der jeweilige Sprachordner eine englische Version der Beispieldatei.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien.

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.

- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.
- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel () wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = “ausgesprochen angemessen” bis 9 = “ausgesprochen unangemessen” zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie () wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = “am meisten bevorzugt” bis 15 = “am wenigsten bevorzugt” zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von “Overall preference” (Allgemein bevorzugt) bis “Snack, with beverage only” (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, “Overall preference” (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.

- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.
- **car_insurance_claims.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **car_sales_uprepared.sav.** Hierbei handelt es sich um eine modifizierte Version der Datei *car_sales.sav*, die keinerlei transformierte Versionen der Felder enthält.
- **carpet.sav.** In einem beliebten Beispiel möchte einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung setzt sich aus drei Faktorenebenen zusammen, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Ebenen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur “Saisonale Zerlegung” berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.

- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.
- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als “AA”, “BB”, “CC”, “DD”, “EE” und “FF” bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customer_information.sav.** Eine hypothetische Datendatei mit Kundenmailingdaten wie Name und Adresse.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.
- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.

- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.
- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dmdata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen über Demografie und Einkäufe für ein Direktmarketing-Unternehmen enthält.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät" . Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.
- **german_credit.sav.** Diese Daten sind aus dem Daten-Set "German credit" im Repository of Machine Learning Databases () an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.

- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell () legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman () verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: “crowds” (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), “audience” (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), “public” (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), “mobs” (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), “primary groups” (Primärgruppen, vertraulich), “secondary groups” (Sekundärgruppen, freiwillig) und “modern community” (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).
- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insurance_claims.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen möchte. Jeder Fall entspricht einem Anspruch.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.

- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs "Quellen" erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.
- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener*(Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen (,) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.

- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.

- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Eine hypothetische Datendatei mit Kauftransaktionsdaten wie Kaufdatum, gekauften Artikeln und Geldbetrag für jede Transaktion.
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln ().
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle

der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.

- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als "good" (gut) "fair" (mittelmäßig) oder "poor" (schlecht) ab.
- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre () vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option "Data Preparation" bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.

- **survey_sample.sav.** Diese hypothetische Datendatei enthält Umfragedaten einschließlich demografischer Daten und verschiedener Meinungskennzahlen.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.
- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei ist eine Untermenge der Datendatei *telco.sav*, allerdings wurde ein Teil der demografischen Datenwerte durch fehlende Werte ersetzt.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das

Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.

- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle () vorgestellt und analysiert.
- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle () vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage (). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder (). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Index

Antwort
  Baummodelle, 70
Ausblenden von Baumverzweigungen, 41
Ausblenden von Knoten
  im Vergleich mit dem Beschneiden, 15

Bäume, 1
  A-priori-Wahrscheinlichkeit, 20
  abhängige metrische Variablen, 83
  Anwenden von Modellen, 83
  Anzahl der Ebenen einschränken, 9
  Auswählen der Fälle in Knoten, 74
  Baum im Tabellenformat, 69
  Baumanzeige skalieren, 44
  Baumanzeige steuern, 26, 46
  Baumausrichtung, 26
  Bauminhalt in einer Tabelle, 26
  Baumstruktur, 43
  bearbeiten, 41
  Bedeutung der Einflussgrößen, 28
  benutzerdefinierte Kosten, 78
  beschneiden, 15
  Bewertung, 83
  CHAID-Aufbaukriterien, 10
  CRT-Methode, 12
  Diagramme, 32
  Effekte der Messniveaus, 53
  Effekte von Wertelabels, 57
  Endknotenstatistik, 28
  Farben, 46
  Farben in Knotendiagrammen, 46
  Fehlende Werte, 23, 93
  Fehlklassifizierungskosten, 17
  Fehlklassifizierungstabelle, 28
  Gewinne für Knoten, Tabelle, 70
  Indexwerte, 28
  Intervalle für metrische unabhängige Variablen, 11
  Knotengröße steuern, 9
  Kreuzvalidierung, 7
  mehrere Knoten auswählen, 41
  mit umfangreichen Bäumen arbeiten, 43
  Modellvariablen speichern, 24
  Modellzusammenfassungstabelle, 67
  Profite, 18
  Regeln erzeugen, 38, 49
  Risikoschätzer, 28
  Risikoschätzer für abhängige metrische Variablen, 88
  Schriftarten, 46
  Speichern vorhergesagter Werte, 73
  Split-Sample-Validierung, 7
  Surrogate, 93, 100
  Textattribute, 46
  Verzweigungen und Knoten ausblenden, 41
  Verzweigungsstatistik ein- und ausblenden, 26
  Werte, 21
Baummodelle, 70
Befehlssyntax
  Auswahl- und Bewertungssyntax für Klassifizierungsbäume erstellen, 38, 49
Beispieldateien
  Lage, 103
Bewertung
  Baummodelle, 83

CHAID, 1
  Bonferroni-Korrektur, 10
  erneut aufgeteilte, zusammengeführte Kategorien, 10
  Intervalle für metrische unabhängige Variablen, 11
  Kriterien für Aufteilen und Zusammenführen, 10
  Maximalzahl der Iterationen, 10
CRT, 1
  beschneiden, 15
  Unreinheitsmaße, 12

Entscheidungbäume beschneiden
  im Vergleich mit dem Ausblenden von Knoten, 15
Entscheidungsbäume , 1
  CHAID-Methode, 1
  CRT-Methode, 1
  erste Variable in Modell aufnehmen lassen, 1
  Exhaustive CHAID-Methode, 1
  Messniveau, 1
  QUEST-Methode, 1, 14

Fehlende Werte
  Bäume, 23
  in Baummodellen, 93
Fehlklassifizierung
  Bäume, 28
  Kosten, 17
  Quoten, 72

Gewichten von Fällen
  nichtganzzahlige Gewichtungen in Entscheidungsbäumen, 1
Gewinndiagramm, 71
Gini, 12

Index
  Baummodelle, 70
Indexdiagramm, 72

Indexwerte
  Bäume, 28

Klassifikationstabelle, 72
Knoten
  mehrere Baumknoten auswählen, 41
Knotennummer
  als Variable in Entscheidungsbäumen speichern, 24
Kosten
  Baummodelle, 78
  Fehlklassifizierung, 17
Kreuzvalidierung
  Bäume, 7

mehrere Baumknoten auswählen, 41
Messniveau
  Entscheidungsbäume, 1
  in Baummodellen, 53
Metrische Variablen
  abhängige Variablen in der Prozedur "Entscheidungsbaum", 83
Modellzusammenfassungstabelle
  Baummodelle, 67

Ordinales Twoing, 12

Profite
  A-priori-Wahrscheinlichkeit, 20
  Bäume, 18, 28

QUEST, 1, 14
  beschneiden, 15

Reduzieren von Baumverzweigungen, 41
Regeln
  Auswahl- und Bewertungssyntax für Klassifizierungsbäume erstellen, 38, 49
Risikoschätzer
  Bäume, 28
  für abhängige kategoriale Variablen, 72
  für abhängige metrische Variablen in der Prozedur "Entscheidungsbaum", 88

Signifikanzniveau für die Aufteilung von Knoten, 14
Split-Sample-Validierung
  Bäume, 7
SQL
  SQL-Syntax für Auswahl und Bewertung erstellen, 38, 49
Startwert für Zufallszahlen
  Entscheidungsbaum-Validierung, 7
Surrogate
  in Baummodellen, 93, 100

Syntax
  Auswahl- und Bewertungssyntax für Klassifizierungsbäume erstellen, 38, 49

Twoing, 12

Unreinheit
  CRT-Bäume, 12

Validierung
  Bäume, 7
Vorhergesagte Wahrscheinlichkeit
  als Variable in Entscheidungsbäumen speichern, 24
Vorhergesagte Werte
  als Variable in Entscheidungsbäumen speichern, 24
  Speichern für Baummodelle, 73

Werte
  Bäume, 21
Wertelabels
  Bäume, 57

Zunahme, 70